# Exploring the Reliability and Validity of Pilot Teacher Ratings in a Large California School District

By

Reino Makkonen

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bruce Fuller, Chair
Professor Xiaoxia Newton
Professor Jesse Rothstein

Fall 2013

Abstract

Exploring the Reliability and Validity of Pilot Teacher Ratings
in a Large California School District

By

Reino Makkonen

Doctor of Philosophy in Education

University of California, Berkeley

Professor Bruce Fuller, Chair

Introduction. Many states and school districts have recently instituted revamped teacher evaluation policies in response to incentives from the federal government as well as a changing political climate favoring holding teachers accountable for the performance of their students. Many of these overhauls have mandated the incorporation of multiple performance indicators — often including rubric-based classroom observation scores, estimated contributions to student test score outcomes, and surveys of students and parents — into teacher evaluations. However, to date there has generally been limited research published on the reliability and validity of this new generation of teacher performance indicators. Much of the existing evidence has come from academic studies rather than from systems that are being implemented in districts for current or eventual professional stakes. Although studies carried out in purely research-oriented settings can provide valuable lessons to the field, they cannot entirely speak to what will be found in real-life implementations.

This three-paper dissertation explores the pilot implementation of a new standards-based multiple-measure teacher evaluation system in a large California school district in 2011/12. It examines both participants' views about the new system (particularly the challenges they faced and the early outcomes they felt were achieved), as well as the reliability and validity of the teacher observation ratings that resulted during pilot implementation. This large California district provides a particularly interesting setting in which to explore these issues, as it has made clear that it is moving forward with the implementation of its new evaluation system, and participating pilot teachers and administrators were well aware that the eventual (and likely near-term) roll out of the reform would include a high-stakes component.

Data and methods. The pilot-year data analyzed in these three papers include transcripts from interviews and focus groups with participating teachers and administrators at five case schools, survey responses from a broader pool of participating pilot teachers and administrators, the disaggregated observation ratings assigned to participating pilot teachers, "value-added" estimates of pilot teachers' contributions to their students' test outcomes during the pilot year, and survey responses from students about their classroom experiences with participating pilot teachers. In addition to describing the extent to which initial implementation of this reform

aligned with the complex logic model that underlies it (paper A), this dissertation also represents an attempt to bring technical measurement methods to bear within the context of a pilot in a large school district, where volunteering teachers and administrators (raters) worked together and tended to know each other (often not the case in academic research settings). Specifically, alongside a more conventional analysis of score distributions and rater agreement, my reliability analysis (paper B) applies the psychometric tools and language of generalizability theory to examine the different sources of variation in pilot teachers' total scores simultaneously, and then I use the relative magnitudes of the variance components to forecast (simulate) reliabilities for differing numbers of observation occasions and/or observer groups.

Given recent shifts toward more grounded argument- and claim-based approaches to validation, paper C adopts a particular integrated line of argument to gather initial evidence about the validity of interpreting participating teachers' classroom observation-based ratings as fair and accurate measures of their pilot-year performance. I first assess the overall face validity of participating teachers' observation ratings (based on teacher survey results), and then present validity evidence relevant to the different parts of the interpretive argument related to those ratings. To do so I initially explore whether significant differences were evident among groups of teachers distinguished by characteristics that should (theoretically) be unrelated to their performance (gender, ethnicity, grade span) and then explore the extent to which teachers' classroom observation-based ratings correlate with other indicators of their pilot-year performance — student survey ratings and value added scores.

Paper A findings. Survey and interview/focus group results indicated that this self-selected group of pilot teachers and administrators generally appreciated the district's new teaching framework and pre/post-observation conferencing process, and participants also tended to report that certain key early outcomes were achieved, including increased reflection by teachers about their performance against the new teaching framework and better understandings of teachers' individual needs for instructional support (although a higher proportion of administrators than teachers reported that this latter outcome was achieved). Time constraints, staffing shortfalls and technology problems were all key challenges cited by both teachers and administrators during the pilot year.

Paper B findings. Analyses of the ratings for the small sample of participating teachers who received a complete set of observational focus element (item) scores from both of their raters across both observation cycles indicated that these teachers tended to be scored higher during the second cycle — although such improvement wasn't universal — and that across cycles the scores from second raters (who typically did not work at the school site) tended to be slightly lower than those awarded by the teachers' supervising site administrator. But ultimately, good agreement was evident between the primary and second raters who scored common teachers. Generalizability analyses indicated that approximately two-thirds of the variation in participating pilot teachers' total scores was attributable to systematic differences among teachers, while the variability associated with the observation cycle (approximately 25 percent) was larger than that associated with rater group (approximately 6 percent). These results were then used to forecast reliability coefficients based on different combinations of rater groups and observed lessons (cycles), and suggested that, based solely on pilot implementation and results from this particular analysis sample, varying the number of observations influenced reliability estimates far more

than varying the number of observers. In particular, as long as at least two observations are conducted, these particular data from this district's pilot year suggest that the reliability of relative distinctions between teachers could remain above the traditional 0.80 threshold with a single rater.

Paper C findings. The group of participating pilot teachers who completed end-of-year surveys generally felt that the observations of their practice conducted during the pilot year represented a valid measure of their effectiveness, and pilot teachers' classroom observation-based ratings were not related to their ethnicity or the grade span they taught (factors that should theoretically be unrelated to performance). Although female pilot teachers did have higher classroom observation-based ratings, on average, than did male pilot teachers, the former group also scored higher than the latter on *all* pilot-year performance measures, suggesting this difference was due to better performance rather than bias. Lastly, low to moderate correlations were evident between pilot teachers' classroom observation-based ratings and their student survey ratings and value-added scores for the 2011/12 year. Thus it's possible that the results from the three piloted measures — arising separately from administrator perceptions, student perceptions, and statistical estimates of teacher contributions to student test outcomes — capture somewhat different dimensions of teacher performance (as has been suggested elsewhere).

Limitations. The uniqueness of this pilot context restricts the generalizability of these findings. The pilot consisted primarily of volunteers, and there was attrition during the pilot year — approximately one-third of the teachers trained in fall 2011 never had any ratings entered online by an observer. In turn, the final pilot sample was comprised of a self selected group of experienced, mostly elementary school teachers who administrators from case study sites tended to characterize as particularly hard working and high performing. Moreover, our research team's limited capacity for qualitative data collection in spring 2012 (we were only able to visit five participating schools) and our low survey response rates (52 percent for teachers and 54 percent for administrators) also limit our ability to generalize findings more broadly. We did not hear the perspectives of those who dropped out of the pilot. Finally, the tools and processes under study were still being revised and fine-tuned by the district during the pilot year; observers were still learning the tools and teachers and administrators were just becoming familiar with the processes and measures. All told, these results likely do not reflect what will be found in any eventual full-scale roll out.

Implications. These results can help inform other districts considering implementing this type of reform. For example, voluntary pilots tend to attract a more experienced and aligned group of teachers and administrators with greater capacity for executing the necessary activities (as was the case here), which in turn would necessitate greater communication and training efforts in subsequent rollouts to non-volunteers. Ultimately, the technical properties of new measures and processes must always be weighed alongside the key contextual aspects of implementation, which can be difficult to predict and, in turn, confront. Not only must practitioners understand the technical tradeoffs involved with particular system design decisions, but researchers cannot ignore the pressing time and resource constraints involved with these types of performance measures. Defining and sticking with a consistent logic model — clarifying purposes and key actors and prioritizing processes and outcomes — can help maintain coherence and forward momentum amidst the political uncertainty common to such reforms.

**Table of Contents**

Paper A: Understanding the implementation of a standards-based multiple measure teacher evaluation reform: Evidence from a pilot year in a large California school district 1

Paper B: Exploring the reliability of teacher observation ratings during pilot implementation in a large California school district 27

Paper C: Exploring the validity of teacher observation ratings during pilot implementation in a large California school district 53

References 83

Appendix: District Framework for Teaching (DFT) schematic for 2011/12 pilot year 91

**Paper A: Understanding the implementation of a standards-based multiple measure teacher evaluation reform: Evidence from a pilot year in a large California school district**

## 1 Introduction

Over the past several years both the U.S. Department of Education and influential organizations like the Bill and Melinda Gates Foundation have funded large-scale efforts to stimulate new thinking and procedures in teacher evaluation and assessment. The federal Teacher Incentive Fund (TIF) program (which began in 2006) bankrolled efforts to develop and implement performance-based teacher compensation systems that "must consider gains in student academic achievement as well as classroom evaluations conducted multiple times during each school year" (USDOE, 2011). In addition, the federal Race to the Top grant competition also contained explicit directives for states and districts to revise their teacher evaluation, compensation, and retention policies to encourage and reward effectiveness (as measured in part by student test scores), and federal guidance on 2010 School Improvement Grants (SIG) released under section 1003(g) of the Elementary and Secondary Education Act stated that any funded district implementing the school "transformation" model option must adopt "rigorous, transparent, and equitable evaluation systems for teachers and principals that take into account data on student growth as a significant factor" (USDOE, 2010: 26). According to recent figures posted on the U.S. Department of Education website, together more than $6.4 billion has been appropriated toward these three programs since 2006. And the Gates Foundation's prominent three-year Measures of Effective Teaching (MET) project dedicated over $45 million to "uncover and develop a set of measures that work together to form a more complete indicator of a teacher's impact on student achievement," enrolling over 3,000 teachers in Charlotte, Denver, Hillsborough County (FL), Memphis, New York City, Dallas and Pittsburgh in the MET effort (Bill and Melinda Gates Foundation, 2011).

Nearly two-thirds of U.S. states have made changes to their teacher-evaluation policies since 2009, and today 25 states require at least an annual evaluation of teachers (Jerald, 2012). Many of these overhauls have increased the number of required observations of teachers' classroom practice and placed more emphasis on standardized test scores. However, there isn't a clear consensus regarding the aims of these reforms, with proponents tending to frame their objective in one of two general ways — either as an opportunity to more efficiently remove teachers perceived as ineffective, or as a means to provide more individualized support to teachers on the job. The former "de-selection" position has been perhaps most famously championed by Stanford economics professor Eric Hanushek, who in 2009 suggested that eliminating 6 to 10 percent of the worst teachers would have a dramatic impact on student achievement, provided the positions were filled with at least "average" teachers (Hanushek, 2009). And as Jerald (2012: 17) emphasized about the alternative (developmental) perspective: "If teachers have been promised one thing from the new evaluations of their effectiveness, it is that they will receive helpful feedback in post-conferences following observations." Among these two competing aims, it is the latter, developmental perspective on evaluation that has been more commonly embraced (at least publicly) by many of the school district leaders and administrators currently

pushing such efforts.[1] Yet systems of individualized teacher feedback and professional development aligned to specified standards are certainly not the norm today, and several initial implementations of multiple-measure teacher evaluation systems in large school districts — including Nashville and Memphis in Tennessee and in Charlotte-Mecklenberg in North Carolina — have been racked by philosophical and logistical problems, including cumbersome and overly-detailed observation protocols, funding shortfalls, and awkward attempts to incorporate math and writing into the curricula for courses not tested statewide (Anderson, 2012; Banchero, 2012). In turn, concerns have been raised that, in the rush to implement new standards-based multiple-measure teacher evaluation systems, policymakers at all levels have left little time for the important trial-and-error process necessary in policy implementation (Mead, Rotherham & Brown, 2012).

At least this much is clear — these are very complex reforms. They often require teachers, administrators, and central office personnel to re-conceptualize their job roles, and they tend to rely on cooperation and "buy-in" from multiple actors and organizations. Given the complexities, challenges in implementation are expected, but have yet to be carefully documented or considered. As Honig (2006: 2) writes, education policy reforms need to be understood in light of "what is implementable and what works for whom, where, when, and why," making more careful study of the implementation of new multiple-measure teacher evaluation systems critical. Such knowledge can benefit an array of actors at various levels of the education system — the district and teacher union leaders wondering about the costs involved with setting up and operating such systems, the site administrators seeking not only clear standards and definitions for assessing instruction but also clarity about the resource supports required to do the work, and finally, the teachers who doubt that such systems can provide them with beneficial feedback and support in addition to potentially fairer professional ratings and job decisions.

At this stage, there are few examples of successful (or unsuccessful) implementations of new standards-based multiple measure teacher evaluation systems to help guide policymakers and practitioners in the creation of these new and highly complex systems and processes. Rather than allowing time for districts and states to learn from each other's successes and mistakes, and to consider the specific elements of district and state context that may allow new teacher evaluation systems to achieve their intended outcomes, policymakers have rushed to implement new evaluation systems. It is yet to be seen if this race to the finish line will result in new systems that do or do not allow districts and states to achieve their purported intended outcomes — to improve teacher practice and thus raise student achievement — without also bringing about myriad unintended consequences.

This paper describes the initial pilot phase of a new standards-based multiple-measure teacher evaluation system in a large California school district in 2011/12. I first present the piloted reform's overall logic model, then highlight several early causal links found to be of key importance during pilot implementation. Specifically, using interview and focus group data gathered from five case study sites and survey responses from participating teachers and

[1] Other influential stages of the professional teaching continuum (for example, preparation and recruitment) have tended to receive less reform focus in many of these recent evaluation-centric policy discussions, although this too is changing (see, for example, Darling-Hammond, 2010).

administrators (who served as teachers' primary observers), I answer three questions related to pilot implementation:

1. During the pilot year, how did participants at case schools perceive the reform and their experiences with its activities?
2. What were the main implementation challenges reported by site administrators, both in case interviews during the pilot year and on end-of-year surveys?
3. To what extent did responding teachers and administrators report on end-of year surveys that pilot implementation achieved two important early outcomes specified in the reform's logic model? Specifically:
    a. Did responding teachers report that they reflected more about their instructional practice than in prior years?
    b. Did surveyed teachers and site administrators agree that pilot activities gave them a better understanding of the areas in which participating teachers could use additional instructional support?

Answers to these questions can yield valuable insights about the early implementation of this complex reform, particularly participants' perspectives on the reform in practice and the type and extent of challenges it presented, as well as whether certain first key steps toward helpful feedback and development (a vital eventual outcome) were achieved in pilot year one. If not, then the complex logic model underlying the reform may break down.

The remainder of this paper proceeds as follows: Section 2 provides background on the district pilot and the logic model that underlies the reform. Section 3 discusses the relatively thin empirical literature that has examined the implementation of standards-based multiple-measure teacher evaluation systems via interviews, focus groups and/or surveys with participants. Section 4 describes the data and methods used to generate findings and Section 5 presents results for each of the research questions outlined above. After a discussion of research limitations (section 6), the paper concludes in section 7 with a discussion of the implications of these results for the district and other districts now starting to implement similar reforms.

## 2 Background

### *System Development and Pilot Context*

The development of the district's new standards-based multiple-measure teacher evaluation system (SBMMTES) came about for a number of reasons. First, the issues with most traditional systems of teacher evaluation also exist in the district. As is the case elsewhere in California, the district's existing teacher evaluation process is based on the Stull Act, a California law that mandates that all school districts establish a uniform system for evaluating certificated personnel. Specifically, the law outlines a specific procedure and timeline for conducting personnel evaluations and stipulates guidelines for the due process rights of those evaluated (The Stull Act, 1971). In the district, the Stull evaluation is guided by an observation checklist, based on the California Standards for the Teaching Profession, that administrators are expected to fill out

when they evaluate teachers, ultimately yielding one of two overall ratings: meets standard performance or below standard performance.[2] A 2009 study found that 99 percent of teachers in the district received a meets standards rating under the district's evaluation system, and that only 64 percent of surveyed teachers reported that evaluation provided them with information and strategies they could use to improve their practice (Teacher Project, 2009). In early 2010, soon after the publication of a final report from the district's teacher performance task force, the local school board directed the district to develop a new system of teacher evaluation and support to address many of the concerns and suggestions raised about the district's current systems.

The district's resulting SBMMTES was developed during the 2010/11 school year, piloted in 2011/12, and was originally communicated to be ready for 2012/13 school year. Upon scale-up, the district's SBMMTES was intended to include multiple measures of teacher effectiveness, including: (1) classroom observations of teacher practice by a site administrator and a second (typically off-site) observer, using protocols aligned with the District Framework for Teaching (or DFT, which was adapted from the Danielson Framework for Teaching);[3] (2) stakeholder feedback surveys of students and parents; and (3) teacher-, grade- and subject-level and school-wide value-added measures of teachers' contribution to student achievement on standardized test scores.

The district spent much of the 2010/11 school year developing the DFT, which was intended to establish a common language and set of norms for effective teaching in the district as well as serve as the eventual foundation for teacher performance reviews and professional development within the SBMMTES. In December 2010, the district convened an ad hoc committee of over 150 teachers, instructional representatives, community partners, and outside consultants to make adjustments to the Danielson Framework for Teaching in order to draft a version appropriate for the district. As part of this process, the district held over 50 focus groups with stakeholders to gather feedback about the potential measure, and reconvened the ad hoc committee in March 2011 to create a final draft of the District Framework for Teaching, which includes five standards, 19 components (2-5 per standard), and 63 elements (2-4 per component).[4] Observation rubrics, templates for the lesson design, teacher self-assessments, and individual growth plans were also developed based on the DFT to be part of the SBMMTES. Also during 2010/11, the district worked with a nationally-recognized research center to generate value-added measures of teacher performance from their students' standardized test scores, and also partnered with local university researchers to develop and field test a set of classroom and school environment surveys for students and parents.

---

[2] The district's Stull form outlines five standards of professional practice on which teachers are rated, including: support for student learning, planning and designing instruction, classroom performance, developing as a professional educator, and punctuality, attendance, and recordkeeping. Each standard has a series of sub-elements that provide greater detail on the specific behaviors that are required to meet the standard. Administrators who evaluate teachers according to the Stull form often receive no or little training on how to observe and assess teachers according to Stull expectations.

[3] The classroom observation measure incorporated teacher self-assessments and lesson planning activities, the actual classroom observations and pre- and post-observation conferences between teachers and observers.

[4] In contrast, the Danielson Framework for Teaching has 4 domains (1 Planning and Preparation, 2 Classroom Environment, 3 Instruction, and 4 Professional Responsibilities) divided into 22 components (that get tracked by observers), and 76 smaller elements. Each component defines a distinct aspect of a domain; two to five elements describe a specific feature of a component (Danielson Group, 2012).

The district began its pilot of the SBMMTES during the 2011/12 school year. The purpose of this pilot year was to learn from the successes and challenges that arose in order to position the system to productively scale-up over time. The pilot field tested the use of teachers' individual growth planning activities, the classroom observation cycle based on the new DFT-aligned protocols, and an online platform for teachers and administrators to report observation notes and ratings. The pilot also field tested stakeholder surveys of students and parents for a subset of the participating teachers. In addition, the district provided school-wide and individual value-added scores to all teachers in the district in 2011/12; only those teachers who taught in statewide-tested subjects received value-added scores.

Overall, the district's pilot was originally scaled to include approximately 100 schools and 700 teachers, who were to be rated against the DFT by their site administrators (primary observers) as well as by a secondary observer (predominantly central office or local/regional office administrators).[5] Participation was voluntary — schools and teacher received stipends for taking part — and all teacher performance reviews had no implications for job tenure or compensation decisions. Yet in part because the district required that every volunteering teacher be in a school with a volunteering principal, only 562 of the initial pool of over 700 teacher volunteers were trained to participate. Another approximately 25 percent of these teachers dropped out after training, for a variety of reasons, including the outspoken opposition of the local teachers union, layoffs, transfers between schools, or evolving concerns over the workload involved. Ultimately, 371 teachers were ultimately rated by any observer in any cycle.[6]

*The System's Logic Model*

The district's new SBMMTES represents a set of tools and processes intended to help administrators and teachers identify areas in which teachers need to improve their practice and then provide them assistance in actually improving in those areas. In this sense, the policy mechanism in use is capacity building, using learning tools to help stakeholders understand how the district defines effective teaching, how teachers rate against this definition and what they can do to improve, and providing resources to help build stakeholders' capacities to translate this knowledge to improvements in practice (McDonnell & Elmore, 1987; Honig, 2006). In addition, when fully implemented, the SBMMTES will also incorporate an element of accountability, in that ineffective teachers who fail to improve will face (yet to be determined) consequences and particularly effective teachers may reap rewards (also yet to be determined) (Loeb & McEwan, 2006; McDonnell & Elmore, 1987). Importantly, however, neither of the system's two overarching theoretical mechanisms for change — professional development in areas where DFT-based ratings indicated that support was needed,[7] or adverse consequences for low

[5] During the pilot, 125 site administrators (principals and assistant principals) served as primary DFT raters, and 210 individuals were trained as second raters, with 167 entering at least one rating for a participating pilot teacher.

[6] An additional 36 teachers started the pilot process by filling out a self-assessment, but were never observed, and the district believes that at least some of these teachers along with another 19 may have been observed during the pilot year, but ratings were never entered for them by a primary or secondary observer. Pilot-year attrition — nearly half of initial teacher volunteers, and approximately a third of those trained, had no ratings entered online by an observer — must be kept in mind when interpreting the results presented here.

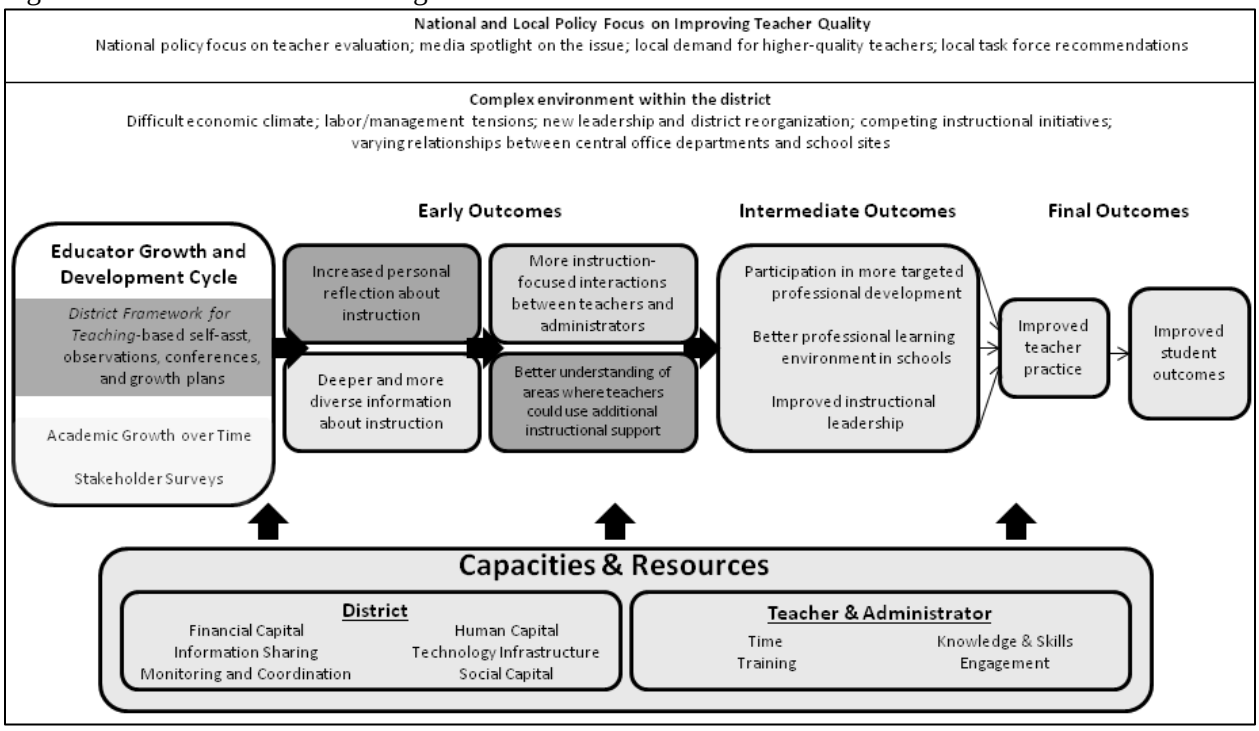
[7] As shown by the lack of dark grey shading in Figure A1, the district did *not* pilot targeted professional development opportunities during the pilot year. The district is currently working on expanding its set of

performance — were effectively enacted during the pilot year, which ended up ultimately representing a trial run for the new measures, tools and processes.

Like any other, the system is intended to reach certain goals, or outcomes, by which its success can be measured. The way that the SBMMTES is intended to reach its outcomes from its beginnings is complicated, and dependent not only on successful implementation of the actual activities that comprise the new work, but also on a host of capacities and resources provided by the district and other important stakeholders, as well as the organizational and political climate within the district that may be instrumental in dictating the eventual efficacy of the policy.

In short, the system's logic model anticipates that the activities required by the process will eventually result in improved teacher practice, and as such, improved student outcomes. Figure A1 displays this multi-step logic model, outlining the paths through which the new activities (left box) — primarily DFT-focused self-assessments, observations and conferences (together designed to increase professional reflection and collaboration around the district's instructional standards and norms), supplemented by information from value-added and stakeholder survey measures where available — are intended to contribute to improved teacher practice and better instructional leadership from administrators, ultimately leading to better student outcomes. District capacity, ongoing training and support, and valid and reliable measures of teacher practice are required to initiate and sustain system activities and results.

Figure A1: District SBMMTES Logic Model




professional development opportunities and aligning them with teachers' needs identified through the new activities and processes. This important support aspect of the full SBMMTES will be included as the district scales up to full implementation.

As noted, though, the 2011/12 pilot year focused only on a subset of system elements for a small proportion of administrators and teachers in the district, and as such was not intended to achieve all of the early, intermediate or final outcomes of the full scale SBMMTES. This paper examines the implementation of just the subset of processes and early outcomes shaded in darker grey in figure A1. Those highlighted in lighter grey indicate those that the pilot intended to test only partially, or that one could not expect to be fully realized (or effectively studied) given the small scale of implementation.[8]

Activities and measures**.** As designed, the district's new SBMMTES is primarily comprised of a series of individual and collaborative tasks and activities that require teachers to engage with the DFT and reflect upon its relationship to their work in the classroom. At the start of the process, teachers rate themselves against the DFT, and this self-assessment is approved by the teacher's supervising administrator. Then, the observation cycle begins. The teacher submits a lesson design and engages in a pre-observation conference with both his or her supervising administrator and a second (off-site) observer, who together then conduct a classroom observation of the teacher's designated lesson, take notes (entered online), and lead a post-observation conference. At the post-observation conference the observers share their notes with the teacher, and then their observation scores for the teacher are entered into the online platform, which the teacher subsequently reviews and acknowledges. Teachers then engage in individual growth planning with their supervising administrators, settling upon a set of development activities for the year (each of which is tied to a particular DFT component or element).[9]

Several key behaviors underlie this work, though. First, teachers must read and understand the DFT, and then they must consider how well they feel they rate against it — for example, how purposively do they group students, or how much critical thinking do their classroom questions inspire? Then teachers are required to engage in a lot of data entry: both self-assessment scores and lesson design. For teachers, the actual observation is fairly traditional, but featuring an administrator in their classroom for the period typing away on a laptop (more online data entry). The conferencing process, in essence, is instructional coaching (rooted in DFT understanding), where the teacher and administrator sit down together and talk about everything entered online and how the evidence relates to the teacher's objectives, both pre- and post-lesson and moving forward. Fundamentally, the bulk of the work involves reading, data entry, and instructional coaching.

[8] For example, since value-added scores and stakeholder surveys were only piloted with a subset of the pilot teacher sample, and since results from these measures were unavailable during the period of our case visits, this paper does not explore the implementation of those measures or the provision of "more diverse" information (e.g., from these new and different sources) about instruction. Instead this paper focuses on the implementation of the DFT-based observation and conferencing process and the initial outcomes (increased professional reflection and improved understandings) I feel are most closely related to it. The relationships between DFT-based observation ratings and scores from the other piloted measures are explored in dissertation paper C.

[9] As the year progresses, the administrator is encouraged to conduct informal observations of the teacher and chart his or her progress on the individual growth plan which may be revised depending on the teachers observed development. The second (spring) observation cycle involves identical activities as cycle 1, but is informed by the data collected throughout the year. (At around this time stakeholder surveys are administered to gather feedback from students and parents, and statewide testing occurs, providing the data for teacher and school value-added scores in the following year.)

Early outcomes. The reform's activities and measures target a particular set of early outcomes. As noted, teachers' completion of the online self-assessment and lesson design are intended to encourage them to authentically *reflect* on their own instructional practice and planning, identify their pedagogical strengths and areas for growth, and critically examine how to build upon their strengths and improve in weaker areas. Such increased personal reflection is considered a first step in encouraging meaningful change in teacher practice, one that ideally should continue throughout the process.

Through the two formal observation cycles and recommended informal reviews, the district's SBMMTES also seeks to increase the number of *instruction-focused interactions* between teachers and administrators and ensure that more discussions about pedagogy and practice occur within schools. Such collaborative interactions are designed to help build a better *understanding of teachers' individual training needs* by relying on the deeper and more nuanced information about practice generated by teachers' reflection as well as by the diverse information provided via the reform's multiple measures. These increased interactions are also designed to build trust between teachers and administrators, help teachers and administrators jointly identify areas for growth (with specific action steps), and increase the perceived legitimacy of both the DFT rubric and the reform's activities among participants. It is assumed that such interactions will be honest, collaborative sessions during which teachers and administrators reflect on evidence and share ideas about the most appropriate improvement strategies.

Together the design's SBMMTES activities and measures are designed to provide teachers and administrators with *deeper and more diverse information* about instruction. Each activity or measure offers a different perspective — a classroom observation may provide a teacher with the knowledge that they need to develop their skills in posing questions to their students, while value-added scores may inform them of the need to improve classroom assessments of student progress. The district assumes that providing teachers and administrators with these diverse sources of information will encourage them to think differently about their practice and better identify areas of strength and challenge for targeted improvements. But as noted, the primary focus of the pilot year was the DFT-based observation and conferencing process, with value-added scores and stakeholder survey results only available for a small subgroup of participating teachers.

Intermediate and final outcomes. As is shown in Figure A1, the successful implementation of the reform is intended to lead to the attainment of the early outcomes described above, which in turn should result in the set of outcomes shown in the box labeled "Intermediate Outcomes." For example, teachers are expected to access *more targeted professional development* through the process, and the more frequent and (hopefully) more meaningful teacher-administrator interactions are intended to help develop a *better professional learning environment* within and across participating school sites. The increased collaboration among participants is also expected to *improve instructional leadership* at school sites (assuming again, a high level of buy-in and engagement from participating teachers and administrators alike).

The ultimate goals of the SBMMTES are *improved teacher practice* and, subsequently, *improved student achievement*. The former is expected to improve through increased personal reflection about instruction, more meaningful and instruction-focused interactions with administrators, and

subsequent participation in better, more targeted professional development. The latter is expected to result directly from the improvements in teacher practice prompted by meaningful engagement in the activities. As teachers improve their practice, it is assumed that students will be exposed to more meaningful learning opportunities yielding direct improvements in their standardized test results.

Key capacities and resources. As in any system-wide reform, a number of specific capacities must exist to enable the successful implementation of the district SBMMTES (see bottom box of Figure 1). Participants must have sufficient *time* to enter the required information online and authentically engage in the observation and conferencing process. If participants don't have enough time — and in the case of administrators, also the staffing resources — to meaningfully participate, then administrators will not be able to effectively judge teachers' strengths and weaknesses and guide them to appropriate assistance, and teachers will not be able to truly reflect on their practice and make necessary changes to their instruction.

Individual participants must also possess a particular set of *knowledge and skills*. Both teachers and administrators must have a modicum of computer skills and a deep understanding of the DFT, and administrators must be certified to observe teacher practice against DFT norms,[10] have sufficient content knowledge to provide meaningful feedback, and the instructional expertise to provide coaching or recommend targeted professional development opportunities.

The logic model described in this section identifies what the district leaders designing the SBMMTES *intend* to happen as the system rolls out. The SBMMTES is truly a process — not fully contained by just the measures and activities included in the individual steps of the cycle. The intended logic model shows that, in fact, the initial steps are only the beginning of the full cycle — it is what the individuals involved in the process *do* with the information provided to them by the measures and processes that will lead to the final intended outcomes. If teachers and administrators do not truly reflect on their instruction and consider how each individual measure fits into a larger picture of teacher proficiencies, and if this reflection does not lead to more instruction-focused and substantive conversations between administrators and teachers, no matter how reliable, valid or accurate the measures are, they will not lead to improved teacher practice. Given these dependencies, the early outcomes matter a lot. If the activities do not cause these early-stage outcomes to occur, the process could break down and will not lead to the intended subsequent outcomes, which are necessary for the district to achieve its final goals of improved teacher practice and student outcomes. This paper seeks to verify empirically whether the earliest aspects of the district's logic model unfolded as intended during the pilot year, by exploring participants' perceptions of the activities they engaged in, the challenges they experienced doing

[10] Discussions with district officials indicate that, as of February 2013, approximately 90 percent of all site administrators who have been trained to observe teachers on the DFT have been certified (10 percent fully certified, 80 percent preliminarily certified), while 10 percent of those trained were not certified. Kane and Staiger (2012) described the certification process for Danielson Framework observers in the MET study: After viewing videos and taking notes, participants' scores had to exactly match at least 50 percent of the scores provided by expert raters and no more than 25 percent of participants' scores could be two or more points off from experts' scores on the four-point scale. Regardless of the certification process used, on end-of-year surveys of pilot participants approximately 40 percent of responding administrators and 30 percent of responding second administrators indicated that they still had questions about how the DFT is used to rate teacher performance.

so, and the degree to which pilot implementation yielded the reflections and understandings that were targeted as key early outcomes.

## 3 Discussion of Related Literature

As noted, this is a complex reform that involves an array of interrelated topics — the complexities of organizational change, education policy implementation, measuring instructional practice, and the ways school principals influence teaching and learning (just to name a few)— and each of these topics has a deep and extensive literature base. A thorough review of all of these interrelated areas of research is beyond the scope of this paper, but several key points do bear mentioning, however, as they have guided my research questions and approach.

First, overcoming longstanding established routines in large organizations like school districts tends to be difficult. The essence of implementation is changing individual behavior (Fixsen et al., 2005), which tends to be hard for policymakers to influence, given their distance from the actual work. Other case study research of large-scale systemic reforms such as this one indicates that actions initiated at some distance from local contexts may constrain local actions, but they do not totally determine them. Instead, contexts shape outcomes; policy effects operate through and within the existing setting and tend to depend on the interpretations and actions of the individuals throughout the system (see, for example, Spillane at el., 2004; Honig, 2006). In the end, implementation plays out via face-to-face interactions among individuals (here teachers and administrators) confronting new activities in concrete social contexts (their schools). Actions that educators take during implementation modify the original policy, suggesting "a co-construction, not as an imposition of policy from the top down nor as a passive flow-through device" (Datnow et al., 1998: 17). At each level, policy implementers negotiate their response and fit their actions to their various demands, priorities, and values (see, for example, McLaughlin, 1987; Resnick, 1991). Not only will the culture of the school mediate educators' actions, but the implementation process will be viewed differently from different perspectives (Datnow et al., 1998).

In recent years, implementations of modified versions of the Danielson Framework for Teaching (FFT) — the focus of this research — have been studied in several other locations, perhaps most notably Cincinnati and Chicago. Cincinnati launched its FFT-based Teacher Evaluation System (TES) in the 2000/01 school year, while in 2008/09 the Chicago Public Schools launched its two-year Excellence in Teaching Pilot, an effort comprised of twice-yearly observations of teachers (again using a modified version of the FFT) as well as pre- and post-observation conferences between the principal and the teacher to discuss evaluation results and teaching practice (similar to the study district's pilot process). Results from interviews, focus groups and surveys in these two contexts yielded interesting teacher and principal perspectives about the implementation of FFT-based processes such as this one. Although in several cases participating teachers reported understanding and accepting as valid the FFT's domains, standards and performance levels (Heneman & Milanowski, 2003; Kimball, 2002), the implementation of such systems imposed significant time constraints and manpower burdens. Specifically, time constraints have been found to inhibit evaluators' ability to provide timely and instructive feedback to teachers (Kimball, 2002; Sartain et al., 2009) and implementation has been found to simply add too many

new demands onto teachers' and administrators' busy workloads (Milanowski, 2001; Heneman & Milanowski, 2003). Chicago's pilot also suffered from weak instructional coaching skills among principals and a lack of engagement among participating principals, which hampered implementation (Sartain et al., 2009, 2011). Here my initial research questions seek to build upon this previous qualitative work, exploring participants' perceptions of the piloted process and what they feel it is supposed to accomplish, as well as the challenges their school sites faced in carrying out the process as designed by the district.

Teachers' reflecting about their pedagogy is an act that many teacher training programs tend to promote (see, for example, Hatton & Smith, 1995), because, as Valli (1997) suggests, such efforts can help individual teachers "break out of unthoughtful habits" (p. 86) and "learn how to recognize weaknesses in their teaching so their students will be more able and motivated to learn" (p. 72). And the meta-cognitive behavior intended by the district's SBMMTES is perhaps best characterized as *technical reflection* (ibid), whereby teachers judge themselves on the basis of external, research-derived criteria (here the DFT) rather than on their own unique situations, inner voices, or ethics. Evidence of increased technical reflection among participating teachers has been found in other recent implementations of the Danielson Framework for Teaching. Heneman and Milanowski (2003), in studying the early implementation of Cincinnati's Teacher Evaluation System (TES), found that participating teachers reported that the evaluation process led them to reflect more on their practice, while teachers in the three FFT-implementing districts studied by Kimball (2002) reported that the new multiple-measure evaluation systems contributed to quality dialogues about instruction between evaluators and teachers. Similarly, in Chicago's recent teacher evaluation pilot, participating principals and teachers both reported that their new Danielson-Framework-based conferences were more evidence-based, reflective, and improvement-focused than those they had engaged in previously (Sartain et al., 2011). Given the study district's emphasis on teachers' technical reflection against the DFT during the pilot year,[11] here I explore whether pilot activities prompted participating teachers to indeed reflect more about their instructional practice than in prior years, as was the case in other recent FFT implementations.

Finally, although research suggests that principals can indeed shape teacher practice (see, for example, Kirby, Berends & Naftel, 2001; Klinger, Ahwee, Pilonieta & Menendez, 2003; Knapp, Copland, & Talbert, 2003), the mechanisms through which principals shape instructional change tend to be indirect, often realized via the principal's influence on the school's culture and professional community (Supovitz, Sirinides & May, 2010). As Ogawa and Bossert (1995) put it, the "context of leadership from an institutional perspective is largely cultural," with administrators affecting "the meanings that other participants fix to organizational events" (p. 229) and engaging them in activities that "shape and reinforce shared values and beliefs, which can produce commitment, or solidarity, leading to coordinated activity" (p. 239).[12]

[11] The pilot process began with all participating teachers self-assessing themselves against the seven DFT focus components, and "Reflecting on Practice" was one of the seven focus components prioritized by the district during the pilot year.

[12] Principals may have less influence on instructional practice in high schools — given those schools' larger and more departmentalized staffs, principals' singular subject matter expertise, and the differentiation among staff roles there — but even in such settings recent research from Chicago suggests that principals can help create a professional climate (via reflective dialogue, teacher socialization, and teacher collaboration) in which teachers

In settings where administrators more directly influence teacher practice — as instructional coaches (the idealized administrator role envisioned by the study district) — research suggests that several key factors matter. In general, the success of instructional coaching requires developing an environment where the roles are clearly defined, the interactions are sustained and substantive (with plenty of opportunities to test understandings, reflect, and debrief), and coaches are knowledgeable, trained, and well matched to the teachers' needs (Fixsen et al., 2005; McCormick & Brennan, 2001; Spillane & Thompson, 1997). Emphasizing this latter point in their recent literature review on teacher professional development, Veen and colleagues (2011: 17) stressed that "it is important to focus on the daily teaching practice, more specifically, the subject content, the subject pedagogical content knowledge and the students' learning processes of a specific subject." Thus my final research question explores a key prerequisite of effective coaching: the extent to which pilot activities gave teachers and administrators a better understanding of the specific areas where teachers could use additional instructional support.

## 4 Data & Methods

The goal of this paper is to describe pilot implementation, specifically the first stages of the reform's logic model (figure A1), and highlight issues that proved particularly important to teachers and site administrators during initial roll out. Data for this paper come from two primary sources. I first draw findings for RQ1 and RQ2 from interviews with principals and focus groups of participating teachers at five pilot schools. This sample of five case study sites was chosen via purposeful, criterion-based selection from the initial pool of approximately 100 pilot schools. Our research team selected case study sites that varied along several key dimensions, including grade level (elementary/middle/high); student demographic makeup; California Academic Performance Index (API) scores; geographic region; and the experience levels and subject areas taught by the participating pilot teachers at the school (see table A1).

Qualitative efforts across these five sites included focus groups and/or interviews with 18 participating EGDC teachers and 11 site administrators, conducted from January to March 2012 (during the pilot year). Although there is no guarantee that these five case schools are typical of all pilot sites, the focus of the qualitative work was on contextualizing the reform and exploring participants' impressions of the process, rather than trying to compare these particular sites to other schools. And indeed, the case studies yielded rich information about participants' perceptions of the DFT and the implementation of pilot activities (at these five sites). All structured interviews and focus groups were recorded and transcribed, and I reviewed the transcripts to identify key themes and patterns of responses for this paper.

---

can succeed, even if the principal cannot direct instructional practice in all subjects (Sebastian & Allensworth, 2012).

Table A1. Descriptive information on case schools

| School | Level | Respondents | | % Minority Category | Geographic Location within District | 2010/11 API Category | Average Years of Teacher Experience |
|---|---|---|---|---|---|---|---|
| | | Teacher | Administrator | | | | |
| School A | Elementary | 2 | 1 | Middle | East | Middle | 10 |
| School B | Elementary | 4 | 1 | Low | North | High | 5 |
| School C | Middle | 4 | 1 | Middle | East | Middle | 7 |
| School D | High | 5 | 3 | Middle | South | Low | 8 |
| School E | High | 3 | 5 | Low | West | Middle | 12 |

Note: To protect the confidentiality of case schools the schools are placed in "percent minority" and API ranges rather than identifying their values directly. These ranges were developed by dividing pilot-eligible schools into quartiles and identifying the bottom quartile as "low minority" or "low API," and the top quartile as "high" minority or API.

A more generalizable analysis of key implementation challenges (RQ2) and early outcomes (RQ3) was afforded by our team's surveying participating teachers and site administrators at the end of the pilot year, building from our case site visits earlier in the spring. Of the 371 participating pilot teachers who received a focus element rating from an observer in 2011/12, 192 completed end-of-year surveys, for a response rate of 52 percent. And among the 125 participating administrators who served as primary observers during the pilot year, 66 submitted end-of-year surveys, for a response rate of 54 percent. These survey responses reflect perceptions of the pilot at the end of the 2011/12 school year, as opposed to case site visits, which as noted took place January through March (amidst pilot implementation).

Table A2 displays descriptive information about the sample of participating pilot teachers and administrators who responded to end-of-year surveys for this project. The 192 responding participating teachers worked in 105 different schools in 2011/12 and generally represent a group of experienced, mostly elementary teachers. During the pilot year, responding participating teachers had taught in their current schools for an average of nine years, in the district for 13 years, and had been teaching for 15 years overall. (California Department of Education data indicate that in 2011/12 the average district teacher had 13.8 years of overall teaching experience.) As shown in table A2, most of the responding participating pilot teachers reported teaching in elementary schools and reported teaching either multiple subjects or English language arts (ELA).

The 66 participating pilot site administrators who responded to the end-of-year survey worked in 58 different (predominantly elementary) schools in 2011/12, and reported having been at their current school site for an average of 5 years, which represented less than half of their average tenure in the district or of their overall administrative experience. The number of pilot teachers reviewed via using the DFT ranged from 1 to 9 across sites, with an average of 3.6 teachers reviewed by participating administrators. Most responding administrators (72 percent) reported reviewing between 1 and 4 teachers using the DFT in the 2011/12 pilot.

Table A2. Descriptive information about respondents from end-of-year surveys*

| | Teachers | Administrators |
|---|---|---|
| Survey respondents (pilot sample) | 192 (371) | 66 (125) |
| Schools represented | 105 | 58 |
| Average experience | Respondent mean | Respondent mean |
| Years in current school | 8.8 | 5.2 |
| Years in district | 13.3 | 11.2 |
| Years as teacher/administrator | 15.1 | 12.5 |
| Grade Level of Current School | % respondents | % respondents |
| Elementary | 51.7** | 51.6** |
| Middle | 26.3 | 23.4 |
| High | 21.9 | 25.0 |
| Subject area (current/previous) | % responses | % responses |
| Elementary-Multiple Subjects | 27.6 | 21.6 |
| English Language Arts | 15.7 | 18.9 |
| Social Studies | 10.6 | 12.4 |
| Science | 10.6 | 6.5 |
| Math | 12.2 | 8.6 |
| Special Education | 4.0 | 11.9 |
| Arts | 5.6 | 11.9 |
| Physical Education | 5.0 | 8.1 |
| Pilot teacher caseload | | % respondents |
| 1 teacher | | 6.2 |
| 2 teachers | | 29.2 |
| 3 teachers | | 21.5 |
| 4 teachers | | 16.9 |
| 5 teachers | | 10.8 |
| 6 teachers | | 7.7 |
| 7 teachers | | 3.1 |
| 8 teachers | | 4.6 |

* The end-of-year survey response rates for the two groups were 52 percent (192 of 371) for participating teachers and 53 percent (66 of 125) for participating school administrators.

** State data for 2011/12 indicate that approximately 65 percent of the district's public schools are elementary schools. (Source: CDE Dataquest: http://dq.cde.ca.gov/dataquest/)

For this paper, survey responses were analyzed using simple descriptive statistics and cross-tabulations to explore potential relationships between various responses and items. Given the relatively low response rates, survey results cannot be generalized to the entire population of participating pilot teachers and administrators. As shown above, results tend to reflect the perspectives of elementary educators (some with ELA backgrounds) and less so the perspectives of secondary educators with, for example, math and/or science backgrounds. There is also clearly a degree of selection bias in our respondent sample. Results indicated that these teachers and administrators tended to believe strongly in the underlying tenets of instructional observation and feedback, and not only did respondents persevere through a pilot process that many considered burdensome, but they also completed an end-of-year survey about it. We did not get to hear the perspectives of teachers who dropped out, and many did. So a certain amount of positive bias should be assumed within this sample, and any negative perceptions might understate actual difficulties.

# 5 Results

*RQ1 During the pilot year, how did participants at case schools perceive the reform and their experiences with its activities?*

Case site discussions with participating pilot teachers and administrators focused on a variety of topics, as our research team sought to get their impressions of the specific piloted activities as well as their perceptions of the DFT content and the overall purpose of the piloted process.

*Pilot Activities*

Opinions varied about the activities that were piloted, with participants generally bemoaning the amount of paperwork and data entry requirements but at the same time expressing appreciation for the observations and conferences (which were now based on some common DFT-based understandings). Regarding the first few steps in the pilot process, five teachers at four case sites and three administrators at three case sites complained that the online self assessment that was required of participating teachers was too long and detailed (and overly time consuming), and respondents felt similarly about the online lesson design template that was required of teachers (which was specifically cited as a challenge by three teachers at two sites and two administrators at two sites). One elementary administrator noted that the length of the self-assessment was counterproductive: "It's almost like you have so much work to do when you're doing that that [teachers] don't really reflect... They're tired by the time they have to reflect." And a participating high school teacher cited the redundancy of the lesson planning requirements. "I did a lesson plan that was 14 pages long. Nobody in this world does a 14-page lesson plan! But then [the district] decided [the form] was too much, so for the second observation, they whittled it down to four pages. But it's still an artificial construct."

Frustrations with the amount of paperwork during the pilot were in part due to problems with the new technology platform/software that was adopted (and revised) during the pilot year. Most of the case site participants we spoke with — 10 administrators and 8 teachers, across all five case sites — expressed some frustration with the software platform/website during the pilot year, alternately referring to it as slow, tedious, repetitive and/or hard to understand, and administrators and teachers both noted that they had to enter information during their evenings and weekends. "There is so much to do [online] that you tend to forget what the whole point is," a middle school math teacher explained. "I lost perspective along the way." His physical education teacher colleague agreed: "I just want to say, 'Forget you guys.' Come and evaluate me with my stupid Stull [form]... I don't have an hour and a day to fill out everything they want me to fill out on that website. I am sorry, I don't. It's just not feasible." (As noted, approximately 25 percent of teachers left the voluntary pilot after training, and at least some of this attrition was due to the workload.)

Despite concerns about the technology and the early steps in the new process, the pilot teachers we spoke with tended to appreciate the pre- and post-observation conferences, with teacher respondents at three case sites welcoming the opportunity to plan things out collaboratively with their administrator and hone in on particular aspects of practice in pre-conferences, while the

benefits of discussing and reflecting on evidence during post-conferences was cited by teachers at four case sites. As two participating teachers explained:

> "*Maybe the software and the tagging and stuff is going to be really too cumbersome, but I hope they don't get rid of the conferencing. I think that having a dialogue about your practice, having observation and dialogue is critical. I don't hear people, teachers or administrators, complaining about the observation or the conferencing. I think those are really valuable... It's forcing a relationship between a teacher and an administrator where there might not be one, a relationship about instruction. That's a good thing*."

> "*The whole idea is to keep working in collaboration with each other. I never got that in Stull, but with the new evaluation system I felt a sense of that collaboration, at least from the level of just having a conversation. For example, one of my administrators asked me the question, for example, 'Well, do you think that the students gained some knowledge when you showed them a second way of solving this math problem?' I thought that was a wonderful question, because instead of just showing it one way, it actually brought out what should've been said, publicly. I thought the conversation made it publicized so that we can all talk about it. There's that part that delves into the teacher's thinking and the lesson planning... That part was awesome*."

In addition, teachers at two case sites and administrators at three sites pointed out that they felt the actual observations under the new system were better, in that they focused on collecting objective evidence on specific aspects of teaching, from the DFT. As a special education teacher emphasized, "You can talk the talk, but when it comes to walking the walk, it needs to be, to me, observed in the classroom itself, and that's really what is happening. Every little piece of evidence is being documented, it's not going to be opinionated. It's going to be factually-based, and you might say things when you're having a conversation to the administrator, but when you do it actually [in the classroom], that's what matters." A middle school administrator agreed: "Being able to look at those descriptors and say, 'This is you, and here is the evidence,' I think is a better reflection for the teacher of what they are doing in the classroom, and it's easier for me to then turn around and say, 'This is where you are at. If you are (rated) Developing, what do you think you can do to change your classroom so that you go up to be Effective?'"

At the same time, teachers at four case sites expressed concerns about the fundamental nature of scheduled, formal observations, pointing out that they may represent inauthentic, rehearsed representations of practice — consistently referred to as "dog and pony shows." Yet the balance being sought is a tricky one to strike, as the (generally appreciated) pre-conferencing and collaborative planning also necessitates a lesson lacking in improvisation.

*DFT content*

Participants generally endorsed the DFT content underlying the piloted process. Four teachers at three case sites and eight administrators at all five sites expressed explicit support for the DFT, generally viewing it as an objective, defined framework that can lead to some shared understandings about effective teaching. "I think the most important thing that (the new process)

did is really create a common understanding amongst administrators of what to look for in a classroom," a high school administrator stated. Agreeing, another noted that the pilot "was a process to refine our (observation) skills as well as the teachers' skills."

Again, the volunteers we spoke with were generally reflective professionals who selected into the pilot process as a way to learn and get better. Nonetheless, teachers did express strong appreciation for the DFT standards and rubric. In particular, three stated:

> "*The observation protocol was better than anything I've experienced in the past, and so that was an enriching experience in and of itself*."
>
> "*It gave me a framework to compare myself with (and an) instructional model goal to aspire to... it makes you want to aspire to that particular standard or substandard of the different aspects of (being) an effective teacher*."
>
> "*I've become much more reflective, and I've used the [DFT] rubric to evaluate myself and actually said, 'I may not be as good as I thought I was.' Because I knew one of the things that I didn't (do well was) hand off the responsibility of the conversation to the classroom, to the students. And so, I think within this past year, I've caught myself dominating the conversation to the point where (now) I catch myself*."

Administrators also generally approved of the DFT content, but four of them at three case sites noted that during the pilot year they felt they were forced to focus on too many aspects of teaching. "Yes, (the DFT) reflects effective teaching, if someone can possibly be adept at all of those things. It's perfection, it's not realistic," one elementary administrator maintained. "There were 63 elements originally. Are you kidding? Even 19 is a lot of different things to observe during one lesson."

Other administrators felt that the extensiveness of the DFT, at least early on, could undermine teachers' potential development. "If you really want someone to grow, you don't tell them, 'These are the 20 things I want to see you growing on,' in what realistically is about seven months of instruction," a middle school principal stated. A high school administrator agreed: "You have to gradually mold and shake [teachers'] behavior and still be respectful. We respect their job that they think they're doing well in the classroom, even though we might know that they could use a little tweaking. So these new things, slapping all this on them and coming up with all these [DFT] variables, it's too much too soon."

*System Purpose*

When asked what they felt the piloted system was supposed to accomplish, five teachers (at two case sites) and three administrators (at two case sites) saw the purpose as basic replacement — that is, exchanging the traditional Stull evaluation process (which they tended to view as perfunctory) with a system based on the DFT's new, more extensive set of standards. "The end game is trying to get an evaluative system that is more rigorous," one high school teacher explained. "I think the intention of the district is to get rid what they perceive to be less effective

teaching (and) to build a progressive discipline procedure where we have the paperwork" to remove teachers perceived as ineffective. (Given that respondents had selected into the process, they tended to agree that such a change was needed.)

At the same time, four teachers at three sites and five administrators at two sites also acknowledged that better support and development of teachers — through reflection around the DFT, more evidence-based observation, and a focus on instructional coaching and feedback — was also a goal. "It's a really different outlook of, everybody has needs. Everybody has areas where they need to grow. You're not proficient in everything, and let's look at some of those areas," one high school administrator stated. "So it's a really different mindset. There's just a fine line between being an evaluator and being a support person."

Concerns about administrators occupying dual roles (as both coach and supervisor) were cited by several site administrators. One high school administrator expressed particularly strong doubts:

> "*I think that the two goals that (the district) has cannot coincide, cannot go together. The eventual goal of who's getting laid off [versus] the other goal of improving teacher's learning. They just don't go together... That's a non-educated point of view of how you can rate an educator. It's too complicated... That premise is the disconnect. I don't know how this (system) is going to tell me what (teachers') PD needs are any more than our current process... We've got over 130 teachers. How (do we) differentiate PD for adults? This isn't going to tell me*."

Site administrators will be the primary implementers of this reform, and their ability to effectively carry out the core work as instructional coaches will influence much of the reform's ultimate success. Respondents we spoke with also emphasized that success also hinges on the ability to eventually connect teachers' areas of need (at least as identified via DFT-based observation and conferencing) with targeted support offerings, a systematic process of connections that has yet to be piloted in the district (and which is now just a work in progress).

As one administrator representative emphasized:

> "*Data is a tool to help teachers and principals, but it is not the be all and end all, and they have more data than they can use now. What are they doing with the data? It's what actually goes on with teachers and kids that matters, and the principal's role is to find the resources and support to improve that, and there are immense challenges right now... (but) if we use that data in the right way, to figure out what supports and training and professional development are needed, and we actually implement that, we will see improvement*."

*RQ2 What were the main implementation challenges reported by participating site administrators, both in case interviews during the pilot year and on end-of-year surveys?*

As noted, several aspects of the piloted system, particularly the DFT content and the conferencing process, were valued by case site participants, but at the same time there were negative aspects of the pilot, including extensive paperwork and problematic technology, that detracted from the new system's perceived legitimacy. This section delves into those latter issues in more detail, augmenting administrators' perceptions of pilot-year challenges cited in case site visits with relevant responses from our broader survey of participating administrators. On surveys, only 30 percent of responding site administrators reported that they agreed with the statement "my school site currently has the capacity to implement activities similar to those in the [pilot] with those teachers who are eligible for evaluation." The key question is then, why? What were the challenges inherent in the pilot that made the idea of bringing it to scale so difficult to imagine? The answer involves three particular and inter-related implementation issues referenced previously, and which can be broadly categorized into *time constraints, technology,* and *staffing resources.*

*Time Constraints*

One of the main resources that teachers and administrators needed to bring to the table in order to make the pilot successful was sufficient time. Teachers and administrators have difficult jobs — especially in the fiscally constrained environment in which the pilot was implemented. During this period the district had laid off administrative and instructional personnel in record numbers due to budget cuts and the economic downturn that hit California particularly hard. The pilot activities replaced traditional Stull activities that had previously been enacted in a relatively cursory fashion by many if not most of the administrators and teachers in the district. In this context, the rigorous pilot activities took more time than what teachers and administrators had previously allocated towards evaluation. It is not particularly surprising, then, that 97 percent of administrators who responded to the project survey (and 93 percent of responding teachers) reported that the pilot activities were too time consuming.

When asked for more specifics about what pilot activities took the most time, administrators cited evidence scoring and tagging, coordination with second observers, and the quantity of focus elements required by the DFT rubric.

Formal Observations and Evidence Tagging. Administrators at all five case sites cited this as a challenge, with one emphasizing that (s)he is tagging evidence online at night and on weekends "because there is no way I could do it during the day. I have two [pilot] teachers, and aside from those two teachers I am evaluating, I want to say, 12 more (via Stull)... There is no way that my (assistant principal) and I could do (all of our) teachers." This sentiment was echoed on end-of-year surveys, with two-thirds of responding pilot administrators reporting that they spent a moderate to great amount of time evidence tagging and scoring, and a full 20 percent reported spending a great amount of time on this activity.

Coordination with second observers. Some of the recent literature on multiple-measure teacher evaluation systems emphasizes the need to rely on two observers in order to generate reliable

observational measures of teacher effectiveness (Doyle & Han, 2012; Ho & Kane, 2013; Kane & Staiger, 2012; Sartain, Stoelinga & Brown, 2011; Taylor & Tyler, 2012). However, the inclusion of second observers in teacher observations during the pilot year appeared to exacerbate the time constraints administrators faced. As two case site administrators explained:

> "*I have two people on site that are doing this, going through the process. Each one has a different second observer. So I have to coordinate. That's another effort that I have to coordinate with two people who are not on site, and who are both secondary people... it's so hard to coordinate people to have that second observer. So, that in and of itself is a huge obstacle... As a logistical problem, it's a nightmare*."

> "*One thing I was confused about in starting this process was when I see (there is) a second observer who's not on our campus... I'm not going to call him in, tell him, 'You be here at such and such a time for observing.' So I figured well... I just did the observation by myself and didn't know that I was really supposed to bring him in... So then I found he really was supposed to come, and that seems an inconvenience... they have so many campuses to go to... he's great to work with, but it was more like I don't really want to bother him with this... And when teachers (are) doing a lesson and the (second observer) isn't available, they're not going to do that lesson the next day*."

This findings was reported more broadly as well, with 58 percent of responding surveyed administrators noting that they faced time difficulties in trying to coordinate with second observers.

Quantity of Focus Elements. The quantity of DFT focus elements (19) on which administrators were expected to rate their teachers in each formal observation cycle contributed to the time constraints noted by administrators during the pilot year. Seventy-four percent of the responding surveyed administrators who reported that they did not collect evidence on all 19 focus elements indicated that they did not do so because they could not collect evidence fast enough during the formal classroom observation periods. Eighty-eight percent reported that the number of elements on which they were expected to collect evidence was not manageable.[13]

As shown in table A3, of the 371 participating teachers rated on at least one focus element by any observer during the pilot year, only 164 (44%) were rated on all 19 focus elements by their primary observer in cycle 1, and only 50 (13%) were rated on all 19 focus elements by both a primary and secondary observer in cycle 1. These figures were even lower in cycle 2, and even lower still for teachers in both cycles. In fact, teachers were only rated on an average of 12.6 and 10.3 focus elements, on average, in cycle 1 and cycle 2 respectively. Of the 125 primary observers, 62 percent rated at least one teacher on all 19 focus elements at least once, but only 31

[13] Eighty-eight percent of administrators also said that they did not collect evidence on all 19 focus elements because there was not evidence about each teaching practice in the lesson(s) they observed. Only a third of administrators reported that they forgot to collect evidence on certain elements and that they did not think certain focus elements were appropriate (32 percent and 35 percent, respectively).

percent rated all of their participating teachers on all 19 focus elements at least once. Even fewer administrators — just 29 percent and 9 percent, respectively — rated one or all of their teachers on all 19 focus elements in both observation cycles (table A3).

Table A3. Focus element (FE) ratings information

| Item | Frequency | Proportion |
|---|---|---|
| Teachers rated on all 19 FE by a primary observer in cycle 1 | 164 | 0.44 |
| Teachers rated on all 19 FE by a primary & second observer in cycle 1 | 50 | 0.13 |
| Average number of FE teachers were rated on in cycle 1 | 12.6 | - |
| Average number of FE teachers were rated on in cycle 2 | 10.3 | - |
| Primary observers rating all their teachers on all 19 FE at least once | 39 | 0.31 |
| Primary observers rating at least one teacher on all 19 FE at least once | 77 | 0.62 |
| Primary observers rating all their teachers on all 19 FE in cycles 1 & 2 | 11 | 0.09 |
| Primary observers rating at least one teacher on all 19 FE in cycles 1 & 2 | 36 | 0.29 |

Note. Teachers were most frequently rated on 4 focus elements, all of which fell under Standard 3: Instruction and include: (1) structures to engage students in learning: standards, based projects, activities, and assignments (71%); (2) using questioning and discussion techniques: student participation (72%); (3) using questioning and discussion techniques: discussion techniques (71%); and (4) using questioning and discussion techniques: quality of purpose and questions (74%).

*Technology Problems*

As noted, administrators at all five case sites cited difficulties with the online platform that was used in the pilot, and technology challenges were also noted within our broader survey sample. The primary challenge highlighted by survey respondents was with entering information from observations online: 45 percent of administrators reported difficulties with technology in completing this activity. This is particularly important because the pilot focused on observers helping teachers set individual goals and tracking their progress towards those goals, all using data entered into the online system.[14]

Some of these time and technology difficulties can be in part traced back to pilot-year trainings. Although surveyed teachers and administrators tended to report high levels of overall satisfaction with their pilot trainings, only 50 percent of respondents reported that the trainings gave them an accurate read of the time commitment involved with the pilot. And satisfaction with training may have impacted implementation; those teachers and administrators reporting higher levels of training satisfaction tended to report fewer implementation difficulties during the pilot ($\rho$=$^{-}$0.32, p<0.01 among teachers and $\rho$=$^{-}$0.45 p<0.01 among administrators, respectively), and administrators who reported greater satisfaction with pilot trainings were also more likely to agree that their school sites had the capacity to scale-up pilot activities with all teachers eligible for evaluation ($\rho$=0.32, p<0.05).

[14] No significant relationship was evident, however, between the number of technology difficulties reported by the administrator and the number of focus element ratings (s)he entered, suggesting that technology challenges did not necessarily impact the number of focus elements on which administrators rated their teachers.

*Staffing Resources*

To carry out this new process, school sites must have the administrative capacity, in terms of sheer staffing numbers, to allow principals and assistant principals to observe and evaluate participating teachers. Clearly, this issue is interconnected with time constraints, as more staff on the ground would allow administrators to allocate less time to operational duties and more time to their new evaluation duties. Administrators at three case sites cited the difficult staffing and scheduling demands that the new system imposed, with survey responses offering similar evidence.

As noted earlier, only 30 percent of responding pilot administrators agreed that their school sites currently have fiscal or staffing resources to implement the piloted system at scale, and at the same time only 26 percent believed that the district had the staffing capacity to bring the new SBMMTES to scale. Beyond staffing, administrators and teachers interviewed at our case sites also expressed concerns that non-participating teachers and administrators would likely have less aligned philosophies and prior experience, which would translate into difficulties down the line. Four administrators and three teachers within our case sites suggested that a lack of sufficient knowledge and skills would hamper some (non-pilot) teachers' abilities to carry out the steps as intended. Specifically, one administrator asserted that,

> "*Nobody has brand new teachers on staff anymore, but if we had brand new, developing teachers, you could counsel them and make them more reflective teachers. But it is impossible to change the practice among the most experienced teachers. They tend to put up a wall and don't want to think anymore. And they truly believe they are doing good work...The impact comes from the depth of reflection involved. Effective teachers will become better, but I don't know that the non-reflective teachers will ever really reflect and internalize this. It won't make the poor teachers better*."

Another administrator put it more bluntly, saying "... (Another non-pilot teacher I evaluate) doesn't have the capacity to fill out these forms. Some teachers don't." Several teachers we spoke with echoed this sentiment, with two stating:

> "*There are a lot of teachers that can't even get their grades into the computer. This (type of work) would be overwhelming for a while*."

> "*The good teachers are going to continue doing well and the poorer teachers, I don't know if this is going to necessarily help them improve or not*."

Compounding several of these implementation challenges was the fact that the district revised its trainings, tools and processes during the pilot year. Administrators at two sites expressed frustration with adjustments in pilot trainings and/or the lack of alignment between the administrator and teacher trainings, the latter of which caused concerns about the shared expectations of administrators and the participating teachers in their schools.

**"***There was a disconnect between what teachers were told in their trainings and what we were told in ours. The principals didn't know what the teachers were hearing…I provided PD on this, but was it aligned to what teachers were doing? No*."

"*…I am not sure how many cohorts of training they did in the summer, but I know that my supervisor, who is my second observer for one of the teachers, she was in the very first cohort. I was in the second cohort. My assistant principal went through the training. I want to say (it was in) September, October which was way, way later and they changed the training and they changed the pieces and what I felt and my supervisor felt was, we thought we were doing the right thing and then we went to a follow up meeting and I called her and I told her, 'Oh, my god. I think we missed all these steps'…So that has been really difficult*."

Additionally, pilot participants at three case sites pointed out that prior field testing or more advance development of support infrastructure would have helped ease implementation and maintain a degree of perceived legitimacy. As two teachers commented:

"*All the new protocols are so complicated and so brand new, and there's so little infrastructure in the district to support it... I'm sure [the district is] learning that it's a real uphill climb. Whether or not that's going to slow things down, I don't know. I don't know if they've learned that*."

The district is "*rolling [the system] out without knowing what they're doing... They've got an idea, but they don't know have the process in place... It makes you feel like the people in charge really don't know what they're doing... they just don't have it together... And the process is just ridiculously involved*."

Of course, learning and improving is a main purpose of a pilot phase, and districts should solidify trainings and details about their pilot programs before full scale up. But here case data caution that such course corrections can also lead to communication gaps and frustration with the process.

*RQ3 To what extent did responding teachers and administrators report on end-of year surveys that pilot implementation achieved two important early outcomes specified in the reform's logic model?*

The district's end goal, as in any new SBMMTES, is to improve teachers' practice, which should subsequently lead to improved student outcomes. But clearly a number of early steps must occur before this can happen, as noted in section 2 of this paper. Specific key early outcomes include increased reflection by teachers about how their instruction aligns with that promoted by the DFT and a better understanding of the areas in which participating teachers could use additional instructional support. In turn, these outcomes should lead to the intermediate and final outcomes posited by the reform's logic model (figure A1). This section assesses whether end-of-pilot-year survey data indicated that the pilot "worked" as intended to achieve these two key early

outcomes. Generally, although the 66 administrator/observers who responded to our survey reported that the pilot was quite successful in achieving these two outcomes, the 192 responding participating pilot teachers were somewhat less enthusiastic, and the qualitative data collected at our five case sites provide a more nuanced view.

A *Did responding teachers report on surveys that they reflected more about their instructional practice than in prior years?*

Survey responses from participating pilot teachers suggested that EGDC was successful in this area: 74 percent of survey respondents agreed that pilot activities "prompted me to reflect more about my instructional practice than in prior years." As further background, the surveyed teachers were also asked if they completed a self-assessment of their "individual needs with an evidence-based process, using a rubric" in the *prior* school year (2010/11). However, those teachers who indicated that, yes, they indeed had prior experience with self assessment were just as likely to agree that the pilot prompted more reflection as those who reported lacking such prior experience. This suggests that the pilot activities did prompt new reflection among participating teachers. There was also no significant relationship evident between responding teachers' pilot-inspired reflection and the number of focus elements on which they self assessed themselves. Those teachers who rated themselves on all 19 focus elements in the pilot year were just as likely to agree that pilot activities prompted more reflection than those who entered fewer self ratings.

However, it is again important to cite the inherent selection bias at play. Survey respondents were motivated volunteers who both participated in the pilot and then also completed a survey about their experience. Outcomes will likely vary among other individuals.

B *Did surveyed teachers and site administrators agree that pilot activities gave them a better understanding of the areas in which participating teachers could use additional instructional support?*

This is a key early outcome, as targeting professional development opportunities requires first a clear understanding of the areas where training(s) might be helpful. On project surveys, 78 percent of responding participating administrators agreed that pilot activities "gave me a better understanding of participating teachers' individual needs." And here again, administrators' responses were not associated with their engaging in similar activities the previous year. Case visit interviews had yielded some earlier insights in this area, with participating administrators at our two case high schools and one middle school each pointing out that the specificity of the DFT helped them identify what to look for in observations and subsequently work to address with teachers. As one of these secondary administrators explained: "I think part of our challenge is going to be looking at the (DFT) and... if there is one or two things to focus on, then turn around and, based on what you know about your teacher, say, 'Work on those two (elements), plus this one, or those two plus these two.' I think that will improve practice."

Surveyed teachers expressed a similar sentiment, but in lower proportions. Fifty-nine percent of responding teachers agreed that pilot activities "gave me a better understanding of my individual needs as a teacher." Prior experience may have influenced teachers' perceptions in this area, though. Those surveyed teachers who reported that the pilot gave them a better understanding of their needs were significantly more likely ($p<.05$) to have engaged in similar activities the previous year. Engaging in similar work the year before may have given these teachers beneficial practice in evaluating themselves; other teachers didn't report the same self-diagnostic benefits from the pilot.

## 6 Limitations of this research

This project's selection and attrition issues have been touched upon already, and they are noteworthy. The teachers and administrators included in the pilot were not randomly selected, and in fact were included in the pilot particularly because of their prior experience with similar processes and/or because of their beliefs in instructional observation and feedback. Principals volunteered (or were asked to volunteer) their schools, and then participating teachers within the schools volunteered or were asked by their principals to participate. Most administrators (65 percent) reported that teachers independently volunteered to participate in the pilot, but 14 percent of administrators reported suggesting that these teachers participate because they represent some of the school's best teachers, and another nine percent suggested these teachers participate because the administrator thought they were particularly well suited to the process. Only five percent of administrators suggested teachers participate because they could benefit from participation. Furthermore, not only was this sample pre-disposed to have greater prior experience and engagement in pilot-like activities, but our respondents also stuck with the reportedly labor-intensive pilot for its full period and then completed surveys and/or interviews about the experience. We did not hear the perspectives of those who dropped out.

In addition, our research team's capacity for qualitative data collection in spring 2012 was somewhat limited, and we were only able to visit five participating schools and interview 11 administrators and 18 teachers (who clearly represent a limited number of participant perspectives). Our low survey response rates (52 percent for teachers and 54 percent for administrators), also limit our ability to generalize findings more broadly.

In sum, the final sample of interview subjects and survey respondents should be considered positively biased toward the process (overstating benefits and understating challenges) and were likely to express opinions that, although perhaps illuminating, do not necessarily reflect those of the full sample of initial pilot volunteers, not to mention the larger population of teachers in the district.

# 7 Discussion

This self-selected group of pilot teachers and administrators generally appreciated the common understandings promoted by the District Framework for Teaching and the pre/post-conferencing process, and tended to report that certain key early outcomes were achieved, although on the latter point a higher proportion of administrators than teachers felt that pilot activities led to better understandings of teachers' individual needs (which may in part be a function of teachers' previous experience with similar activities).

Time constraints, staffing shortfalls and technology problems were key (related) challenges for both teachers and administrators during the pilot year, with administrators citing particular challenges with online evidence tagging and scoring (due in part to the number of focus elements), while teachers primarily bemoaned the time burden imposed by the online lesson design and self-assessment. But here trainings may serve as an important leverage point, as our data suggest that relevant, consistent training — targeted at such key areas as the DFT, the online platform, and instructional coaching techniques (with trainings ideally delivered separately for elementary and secondary educators) — might help reduce confusion and frustration that was experienced during the pilot year and increase participant support, buy in, and fidelity to the systems' planned implementation.

Although plans for the district's SBMMTES have shifted since the beginning of the pilot year and a degree of uncertainty remains, other districts and states embarking upon (or considering) similar reforms can learn from this district's experience. Voluntary pilots may attract a more experienced and aligned group of teachers and administrators with greater capacity for executing the necessary activities, as was the case here, which in turn necessitates far greater communication and training efforts in subsequent rollouts to non-volunteers. Ultimately, the technical properties of new measures and processes must always be weighed alongside the key contextual aspects of implementation, which can be difficult to predict and, in turn, confront. Defining and sticking with a consistent logic model — clarifying purposes and key actors and prioritizing processes and outcomes — can help maintain coherence and forward momentum amidst the political uncertainty common to such reforms.

## Paper B: Exploring the reliability of teacher observation ratings during pilot implementation in a large California school district

### 1 Introduction

As education policymakers increasingly turn their sights to evaluating teacher effectiveness in response to federal initiatives like the Teacher Incentive Fund and Race to the Top grant programs, the Danielson Framework for Teaching (FFT), originally developed in 1996, is receiving a tremendous amount of attention by education leaders looking for an all-encompassing, cross-subject framework to evaluate their teachers.[15] After several years of planning and preparation, in fall 2011 a large urban California school district began piloting a new process of two-observer classroom observations with pre- and post-conferences, as part of its new multiple-measure system of teacher evaluation. This observation and conferencing process is based on the district's new District Framework for Teaching (DFT), an adapted version of the Danielson Framework for Teaching, positioning the study district alongside other large districts like Pittsburgh, Cincinnati, and Hillsborough County (Tampa FL) and state education agencies like Illinois, New Jersey and New York that are now adopting or modifying the FFT for use in their systems of teacher evaluation.

Within the growing research base on the properties of, and results from, new observational instruments like the FFT, this paper examines the consistency of teachers' ratings across two observation occasions (lessons) and across two different raters — the teacher's supervising administrator and a second observer, typically an administrator based off-site — during the pilot phase of this district's new standards-based, multiple-measure teacher evaluation system (SBMMTES). Alongside a more conventional analysis of score distributions and rater agreement, this paper also applies the psychometric tools and language of generalizability (G) theory to examine the different sources of variation in pilot teachers' total scores simultaneously, apportioning variance to differences between teachers, differences in raters, and changes over observation occasions (lessons).[16]

This type of work is particularly important today, amid growing consensus that teachers are the most important school-based influence on student achievement — research suggests that having a particularly good teacher will positively affect students' current academic performance as well as their future success (Rockoff, 2004; Rivkin, Hanushek, & Kain, 2005; Goldhaber et al, 1999;

[15] Originally published in a 1996 book by Charlotte Danielson (*Enhancing Professional Practice: A Framework for Teaching*), the Framework for Teaching is a set of components of classroom instruction that are "grounded in a constructivist view of learning and teaching" (Danielson Group, 2012). Burgeoning efforts are also underway advancing the development of measures of subject-specific content knowledge and pedagogy, including, for example, the Protocol for Language Arts Teaching Observations (or PLATO, developed by Dr. Pam Grossman at Stanford University) and the Mathematical Quality of Instruction protocol (or MQI, developed by Dr. Heather Hill of Harvard University). Such efforts are not the focus in the study district or in this paper, however.

[16] Similar to the reliability analyses conducted as part of the recently-concluded Measures of Effective Teaching (MET) project (Kane & Staiger, 2012; Ho & Kane, 2013), the generalizability portion of the paper seeks to separate pilot teachers' overall "true" or "universe" ratings from variations due to raters or the occasion (lessons) observed. Although targeted, within-year in-service professional development will be part of the study district's SBMMTES eventually — as teacher growth and development during the evaluation year is a priority of the program's designers — such opportunities were not part of the pilot year in 2011/12. Thus any within-year improvements in 2011/12 were likely due to individual learnings rather than any systematic development effort.

Chetty, Friedman, & Rockoff, 2012). At the same time as researchers have been proving what many parents and school-based practitioners have long-known, research has also highlighted the major flaws in current systems of teacher evaluation, with reformers tending to frame their objectives in one of two general ways: either as an opportunity to more efficiently remove teachers perceived as ineffective, or as a means to provide more individualized support to teachers on the job. Classroom observations are receiving increased focus because they are generally considered the most direct way to measure instructional quality (Clare, 2000; Newton, 2010), yet at the same time, reliable scoring of teachers' observed classroom performance is a complex endeavor. Ratings result from interactions between the observational rubric, the observers using it, the training of those observers, and the scoring system used, and each of these design elements can impact reliability (Hill, Charalambous & Kraft, 2012). Clearly, both teachers and other users of evaluation ratings — including both those seeking efficient removal or more targeted professional development — want to be sure that the score is mostly due to observed practice rather than raters' idiosyncratic views about what good practice looks like, or the time of year or occasion (lesson) when the teacher was observed.

This paper represents an attempt to bring technical measurement methods to bear upon this key policy topic, within the context of an initial implementation in a large urban school district. Unlike the Gates Foundation's large scale Measures of Effective Teaching (MET) research project that just concluded its work across several large school districts, this study was not part of a controlled academic research experiment with videotaped lessons and anonymous expert observers. Instead, results arose from the actual implementation of a new (pilot) policy in schools where volunteering teachers and administrators worked together and tended to know each other. This makes this work particularly relevant from a policy perspective. Studies carried out in purely research-oriented settings can provide valuable lessons to the field, but they cannot entirely speak to what will be found in real-life implementations for current or eventual stakes.

Results in this paper indicate that, for the volunteering teachers with scores across observers and cycles, inter-rater agreement was generally high (over 0.7), with most of the variability in teacher's total observation scores attributable to systematic differences between participating teachers and changes between observation cycles. This (relatively) high reliability is notable given the myriad political, contextual and capacity realities that tend to face the large districts now implementing such systems. In such environments, principals and teachers are faced with various duties in addition to participating in observations, and as mentioned above, participants may not be as highly trained or detached as in solely research-oriented studies. In addition, the generalizability study results here indicate that — for this particular analysis sample of teachers observed via the process implemented during the pilot year (i.e., with some observer collaboration) — the number of observations influenced score variation more than did the number of observers, suggesting that the reliability of relative DFT-based distinctions between teachers could potentially remain at or above traditional benchmark levels with a single rater conducting multiple observations. These results can provide some guidance to districts that face important tradeoffs between the technical properties of a measure — with some recent literature (for example, Ho & Kane, 2013) emphasizing the need to rely on multiple observers in order to generate reliable observational measures — and other operational considerations, such as allowing principals to maintain a manageable workload and providing teachers enough time to fulfill all of their student- and parent-oriented duties. In the end, districts must judge competing

considerations and strike a balance, and this district's context allows for an important analysis of observation scores implemented in a "real world" setting.

The remainder of the paper proceeds as follows: Section 2 provides background on the district pilot context, and section 3 provides a brief review of the limited research on the reliability of FFT-based observational measures. Section 4 presents the paper's research questions, and sections 5 and 6 describe the data and methods used. Section 7 describes the paper's results, and the limitations of this research are discussed in section 8. The paper concludes in section 9 with a discussion of the policy implications of this work.

## 2 Background

The study district's new multiple-measure teacher evaluation system arose for a number of reasons. First, the shortfalls within most conventional systems of teacher evaluation also exist in the district. An study of the district in 2009 found that 99 percent of teachers received a meets standards rating, yet only 64 percent of teachers surveyed in the study reported that the evaluation provides them with information and strategies they could use to improve their practice, and only 64 percent of principals reported that the evaluation system allowed them to adequately address issues of poor performance among their faculty (Teacher Project, 2009).

In early 2010, soon after the publication of the final report from the district's teacher performance task force, the local school board directed the district to develop a new system of teacher evaluation and support to address many of the concerns and suggestions raised about the district's current systems. In November 2010, the district convened an ad hoc committee of internal and external partners, including over 150 teachers, instructional officials and outside consultants to craft the District Framework for Teaching (DFT), which as noted represents a modified version of the FFT, with five standards (1 Planning and Preparation, 2 Classroom Environment, 3 Instruction, 4 Additional Professional Responsibilities, and 5 Professional Growth), 19 components (2-5 per standard), and 63 elements (2-4 per component).[17] The DFT is intended to establish a common language and set of norms for effective teaching in the district, as well as to serve as the eventual foundation for teacher performance reviews and professional development within the district's SBMMTES. Observation rubrics, templates for the lesson design, teacher self-reviews, and individual growth plans were also developed based on the DFT to be part of the district's new standards-based, multiple-measure teacher evaluation system (SBMMTES).

In the 2011/12 school year, the district began piloting the SBMMTES with approximately 400 teachers and approximately 125 site administrators across 100 schools within the district. Participation by these teachers and administrators was voluntary, and each school and teacher received a stipend for taking part. All teacher performance reviews had no professional stakes

[17] In contrast, the Danielson Framework for Teaching has 4 domains (1 Planning and Preparation, 2 Classroom Environment, 3 Instruction, and 4 Professional Responsibilities) divided into 22 components (that get tracked by observers), and 76 smaller elements. Each component defines a distinct aspect of a domain; two to five elements describe a specific feature of a component (Danielson Group, 2012).

attached to them during the pilot year. The purpose was to learn from the successes and challenges that arose and make adjustments in order to position the system to productively scale-up over time. There was some attrition during the pilot year, as approximately one-third of the teachers trained in fall 2011 did not end up having a rating entered online by an observer (see section 5).

The pilot initiated the use of teachers' individual growth planning activities, the classroom observation cycle based on the new DFT-aligned protocols (with scores representing the focus of this paper), and the online platform for teachers and administrators to report observation notes and ratings. The pilot also included stakeholder surveys of students and parents for a subset of participating pilot teachers.[18] In addition, the district provided school-wide and individual value-added scores to all teachers in the district in the 2011/12 school year (only those teachers who taught in tested subjects received value-added scores). These latter two measures are not the focus of this paper, however.

## 3 Discussion of Related Literature

The term *reliability* is used to describe the overall consistency of a measure — a measure is said to have a high reliability if it repeatedly produces similar results under consistent testing conditions (AERA, APA, NCME, 1966, 1999). Three classes of reliability estimates tend to be the most well known. *Inter-rater reliability* (also referred to as rater agreement) defines the degree to which test scores are consistent when measurements are taken by different people using the same methods. Alternatively, *test-retest reliability* defines the degree to which test scores are consistent from one test administration to the next, while in other instances the primary focus is on the *internal consistency* of results across items within a particular test or instrument. Estimates of reliability generally range from 0 to 1, with higher values indicating a more stable measure of the ability of interest; for example, DeVellis (1991: 85) characterized values between 0.7 and 0.8 as "respectable" and values above 0.8 as "very good."

To date there have been relatively few empirical studies of the results from the current generation of teacher observation protocols (Bell et al., 2012; McREL, 2012), and only a limited number of articles have yet been published on the reliability of the Danielson Framework for Teaching. Sartain and colleagues (2011) examined whether the FFT-based ratings provided by principals in Chicago's Excellence in Teaching Pilot agreed with ratings provided for the same lesson by highly trained external observers who did not know the teachers. Principals were more likely to rate a teacher at the highest level on the four-point scale than were external observers, who tended to rate those same teachers at the second-highest level.[19] Analyzing rater agreement

[18] The stakeholder (parent and student) surveys piloted for a subset of participating teachers in 2011/12 were administered towards the close of the school year. Preliminary data provided to the research team indicated the district received at least one completed parent survey for 18 different elementary teachers and for 23 secondary teachers (7% response rate among parents), and from students the district received at least one completed classroom experience survey for 73 different elementary teachers and 365 different secondary teachers (41% response rate among students).

[19] This does not imply that either observer was "right," however. Sartain and colleagues (2011) explained that, while principals used the FFT's Distinguished rating more often, these Distinguished teachers had higher value-added

for Cincinnati's Teacher Evaluation System (TES) — a rigorous process based on a modified version of the FFT and featuring extensive observer training and four observations during the evaluation year (three by an external peer evaluator and once by a site administrator) — Milanowski (2011) found a high level of average absolute agreement between the two observers rating teachers on different occasions on TES Domain 2: Creating an Environment for Learning (73 percent rater agreement) and TES Domain 3: Teaching for Learning (79 percent rater agreement).

The Gates Foundation's large-scale MET project examined the reliability of the FFT in two recent studies (Kane & Staiger, 2012; Ho & Kane, 2013). In their generalizability analysis to examine the relative size of different sources of variation in FFT-based ratings, Kane and Staiger (2012) found that residual (unobserved) sources of variation, including disagreements among raters that varied by lesson or class section, accounted for 43 percent of the variance in ratings. With this information, the authors reported that the rating of one lesson by one rater produced a relatively unreliable assessment of teacher effectiveness (a reliability of only 0.37), while having four lessons each scored by a different rater yielded much higher reliability (0.67). In their similar psychometric analysis, analyzing ratings from FFT domains 2 (Classroom Environment) and 3 (Instruction) in Hillsborough County (FL), Ho and Kane (2013) emphasized the need for multiple observers, finding that administrators rated their own teachers 1/3 of a standard deviation higher on average than external administrators and 2/3 of a standard deviation higher than peer observers — although these discrepancies had little influence on teachers' relative rankings. A single observation by a single rater had low reliability (between .27 and .45 in the study), and adding raters improved reliability more than adding observations by the same rater (Ho & Kane, 2013). As noted, though, these MET results tended to be derived from contexts where outside observers were reviewing videotaped lessons, and not in situations where the schools and districts were actually implementing new systems for current or eventual stakes, as is the case in this district.

Together, these studies have identified different sources of variability in FFT-based observation scores, and researchers are just now beginning to examine such systems across varied contexts.[20]

---

measures than the teachers the principals rated as Proficient, suggesting to the authors "either that principals are correctly identifying Distinguished practice or that they used historical knowledge of the teacher (unknown to the observer and outside of the evidence of the classroom observation) to form a better picture of teacher effectiveness" (p. 16). When Sartain and colleagues accounted for teachers' previous evaluation ratings in their statistical models, much of the variation between principal and observer ratings disappeared.

[20] For example, two large districts participating in the MET project, Pittsburgh and Cincinnati, structure their Danielson FFT-based observation measures and processes slightly differently from one another (and from the study district). Specifically, Pittsburgh's Research-Based Inclusive System of Evaluation (RISE) rubric is organized around the four Danielson-based domains but has 24 sub-components, 12 of which represent "power components" that receive additional focus from evaluators during classroom observations (conducted at least twice per year) and which end up comprising teachers' summative (end-of-year) rating (see http://www.pps.k12.pa.us/cms/lib07/PA01000449/Centricity/domain/30/document%20library/RISE_rubric_2011-12.pdf). In the Cincinnati TES, site administrators and peer reviewers rate teachers (via classroom observation and portfolio review) on 15 standards within four Danielson FFT-based domains. All Cincinnati teachers receive formative, single-observation reviews annually, with more comprehensive (four-observation) reviews scheduled at defined intervals — the teacher's first year as a new hire, his or her fourth year, and then every five years after that point. However, a presentation by Cincinnati officials at a recent National Center for Teacher Effectiveness (NCTE) conference suggested that the district has had some challenges related to the consistency and interpretation of evidence collected (see http://www.gse.harvard.edu/ncte/news/NCTE_Conference_Cincinnatti.pdf).

Moreover, developers of observational instruments have traditionally paid little attention to how decisions about the design of the observational system (such as the number of raters and observations involved), will affect the reliability of the resulting teacher scores (Hill, Charalambous & Kraft, 2012).

## 4 Research Questions

Given the limited research to date on the reliability of teachers' Danielson FFT-based ratings as implemented in practice, this study analyzes observation data on participating pilot teachers to explore the variability in their scores and offer relevant policy insights about system design. Specifically, I ask:

1. How consistent across observation cycles and raters were the DFT-based scores assigned to pilot teachers?
2. How much of the variance in pilot teachers' total DFT scores was attributable to systematic differences between teachers?
3. Using these pilot year results, how might a different number of observers or observation cycles influence the reliability of teachers' total DFT scores?

## 5 Data

The district's DFT-based rubric for rating teacher performance includes four rating levels (Ineffective, Developing, Effective and Highly Effective), with descriptors defining performance on each element at each level. At the start of the 2011/12 pilot year, the district defined seven "focus" components and 19 "focus" elements upon which participating teachers were to be assessed. The seven focus components were:

1d Planning & Preparation: Designing coherent instruction (contains four focus elements [FE])
1e Planning & Preparation: Designing student assessment (one FE)
2b Classroom Environment: Establishing a culture for learning (two FE)
3b Instruction: Using questioning and discussion techniques (3 FE)
3c Instruction: Structures to engage students in learning (3 FE)
3d Instruction: Using assessment in instruction to advance student learning (4 FE)
5a Professional Growth: Reflecting on practice (2 FE)

Participating pilot teachers were to be assigned one of the four performance ratings on each of the 19 DFT focus elements in two separate lesson observations (observation cycle 1, from October 2011 to February 2012, and then again in observation cycle 2, from March to May 2012) by each of their two raters — the teacher's supervising site administrator (primary rater) and his

or her second rater (typically based off site).[21] The district initially recruited approximately 700 volunteer teachers to participate in the pilot. However, in part because the district required that every volunteering teacher be in a school with a volunteering principal, only 562 of the initial teacher volunteers were trained to participate. Another approximately 25 percent of these teachers dropped out after training, for a variety of reasons, including the outspoken opposition of the local teachers union, layoffs, transfers between schools, or evolving concerns over the workload involved. Ultimately, during the pilot year 371 teachers were rated by any observer in any cycle.[22] In total, 125 site administrators (principals and assistant principals) served as primary raters during the pilot year, and 210 individuals were trained as second raters, with 167 entering a rating for a participating pilot teacher. Table B1 offers descriptive information about these raters, compiled from surveys of these two rater groups at the end of the pilot year.

Table B1. Descriptive information about primary and external raters from end-of-year surveys*

| | Primary Raters | Second Raters |
|---|---|---|
| Job title | % respondents | % respondents |
| Principal | 77.3 | 5.6 |
| Assistant Principal | 15.2 | 3.8 |
| Central Office Staff | 3.0 | 16.5 |
| Local/Regional Office Staff | 4.5 | 74.7 |
| Grade span previously taught | % responses | % responses |
| Elementary | 48.9 | 53.7 |
| Middle | 29.5 | 26.3 |
| High | 21.6 | 20.0 |
| Subject area previously taught | % responses | % responses |
| Elementary-Multiple Subjects | 21.6 | 54.9 |
| English Language Arts | 18.9 | 16.5 |
| Social Studies | 12.4 | 4.4 |
| Science | 6.5 | 4.4 |
| Math | 8.6 | 4.4 |
| Special Education | 11.9 | 3.3 |
| Arts | 11.9 | 8.8 |
| Physical Education | 8.1 | 3.3 |
| Pilot teacher caseload | % respondents | % respondents |
| 1 teacher | 6.2 | 31.3 |
| 2 teachers | 29.2 | 45.0 |
| 3 teachers | 21.5 | 13.8 |
| 4 teachers | 16.9 | 6.3 |
| 5 teachers | 10.8 | 3.8 |
| 6 teachers | 7.7 | – |
| 7 teachers | 3.1 | – |
| 8 teachers | 4.6 | – |

* End-of-year survey response rates for the two rater groups were 53 percent (66 of 125) for participating primary raters and 49 percent (82 of 167) for participating external raters.

[21] The district did not specifically define at the start of the pilot year whether primary and external observers were to input their ratings independently or collaboratively.

[22] An additional 36 teachers started the pilot process by filling out a self-assessment, but were never observed, and the district believes that at least some of these teachers along with another 19 may have been observed during the pilot year, but ratings were never entered for them by a primary or secondary observer. Pilot-year attrition — nearly half of initial teacher volunteers, and approximately a third of those trained, had no ratings entered online by an observer — must be kept in mind when interpreting the results presented here.

During the summer and fall of 2011, multiple cohorts of potential observers participated in a 32-hour, weeklong training on the DFT. Observers watched and took notes on videos, then tagged evidence from their notes and applied ratings to their observations. Tagged evidence and ratings were reviewed by training leaders and rated on an evidence rubric and scoring accuracy measures. Based on these results, the observer was deemed either "certified" or "preliminarily certified," and some were provided with additional support before attempting certification again. However, not all observers in the pilot were certified or preliminarily certified (approximately 10% didn't attain certification), and the district has continued to work with them towards certification.[23]

Table B2 displays the DFT rubric language associated with the two most common DFT focus elements rated during the pilot year, both of which fall under component 3b (Using questioning and discussion techniques): Quality and Purpose of Questions (73 percent of participating teachers were scored on this element) and Student Participation (72 percent). As shown, Highly Effective performance in this area involves practices that engage and challenge students to apply reasoning and monitor their discussion, while Ineffective performance involves closed questioning dominated by a limited number of voices.

Table B2. Performance descriptions for 2 most common focus elements rated during pilot year

| Standard 3: Instruction | | | | |
|---|---|---|---|---|
| Component 3b: Using questioning and discussion techniques | | | | |
| Focus Element | Ineffective | Developing | Effective | Highly Effective |
| 3b1. Quality and Purpose of Questions: Questions are designed to challenge students and elicit high-level thinking | Teacher's questions are largely closed in nature or not relevant. Questions do not invite a thoughtful response. Questions do not reveal student understanding about the content/concept, or are not comprehensible to most students. | Teacher's questions are a combination of open and closed questions of both high and low quality, or delivered in rapid succession. Only some questions invite a thoughtful response that reveals student understanding about the content/concept under discussion. Teacher differentiates questions to make them comprehensible for most students. | Teacher's questions are mostly open-ended and require student thinking. Most questions invite and reveal student understanding about the content/concept under discussion. Teacher differentiates questions to make learning comprehensible for groups of students. | Teacher's questions challenge students to think and invite students to demonstrate understanding through reasoning. Students themselves formulate many questions to advance their understanding. Teacher differentiates questions to make learning comprehensible for all levels of English learners and special needs students in the class. |

[23] Discussions with district officials indicate that, as of February 2013, approximately 90 percent of all site administrators who have been trained to observe teachers on the DFT have been certified (10 percent fully certified, 80 percent preliminarily certified), while 10 percent of those trained were not certified. Kane and Staiger (2012) described the certification process for Danielson FFT observers in the MET study: After viewing videos and taking notes, participants' scores had to exactly match at least 50 percent of the scores provided by expert raters and no more than 25 percent of participants' scores could be two or more points off from experts' scores on the four-point scale. Regardless of the certification process used, on end-of-year surveys of pilot participants approximately 40 percent of responding administrators and 30 percent of responding second administrators indicated that they still had questions about how the DFT is used to rate teacher performance.

| | | | | |
|---|---|---|---|---|
| 3b3. Student Participation: Techniques are used to ensure all students participate in discussions | The teacher and/or a few students dominate the discussion. | Teacher inconsistently engages all students in the discussion, but instructional and questioning techniques result in only limited success. | Teacher attempts gradual release from teacher-directed to student-initiated participation. All students participate when coached by teacher. | Teacher functions as facilitator using instructional and questioning techniques that engage all students in discussion. Students themselves ensure that all voices and ideas are heard in the discussion. |

The de-identified observation data analyzed for this paper included, for each of the 19 DFT focus elements assessed during the pilot, the teacher's rating (coded 1-4), the observation cycle (coded 1-2), and the rater who provided the score (primary or second). During the first observation cycle (October 2011 to February 2012), 164 participating pilot teachers were scored on all 19 focus elements by their primary rater, while 50 of these teachers also had a comprehensive set of scores entered by their second rater (table B3). These 50 teachers received focus element scores from 24 different primary raters and 32 different second raters, with most of these observers (75 percent of primary raters and 84 percent of second raters) only rating one or two teachers.

During the second observation cycle (March to May 2012), a different group of 126 pilot teachers were scored on all 19 focus elements by their primary rater, and 33 of these teachers were also scored online across all focus elements by their second rater during cycle 2. Fifteen different primary raters and 19 second raters rated these 33 teachers (with most observers again rating either one or two teachers).

Participants and system developers learned about the process and made adjustments during the pilot year, and not all participating teachers ended up being scored by both raters in both cycles, nor were all teachers rated on all seven focus components and all 19 focus elements in each cycle.[24] Eighty-eight participating pilot teachers were scored on all 19 focus elements by their *primary* rater in *both* cycles, and 16 teachers had a complete set of 19 focus element scores from *both* raters in *both* observation cycles (see top set of rows in table B3).

[24] Among the subset of surveyed administrators and external raters who responded that they did *not* collect evidence on all 19 focus elements during the pilot year, 88 percent reported that they did not do so because they only witnessed evidence of certain teaching practices in the lesson observed. And although the district established the two-rater system for the pilot year, the district initially allowed observers some discretion in how primary and second observers input ratings (at the end of the pilot year the district asked observers to agree on a single set of ratings). On end-of-year surveys, however, 64 percent of all responding observers (primary and second) stated that they entered teacher scores separately in cycle 2, while 36 percent indicated that they collaborated and entered a single set of agreed-upon scores for the teacher in cycle 2. (The end-of-year survey response rates among these two observer groups were 51 percent for participating pilot site administrators [64 of 125] and 45 percent for participating second observers [75 of 167].)

Table B3. Participating pilot teachers rated on focus elements by observers across cycles

| 19 focus element ratings | Cycle 1 | Cycle 2 | Both Cycles |
|---|---|---|---|
| a. From Primary Rater | 164 | 126 | 88 |
| b. From External Rater | 76 | 61 | 36 |
| c. From Both Raters | 50[a] | 33[b] | 16[c] |
| 10 Standard 3: Instruction ratings | Cycle 1 | Cycle 2 | Both Cycles |
| a. From Primary Rater | 216 | 159 | 124 |
| b. From External Rater | 159 | 84 | 51 |
| c. From Both Raters | 81 | 47 | 27[d] |

[a] Within this sample of 50 fully-rated cycle 1 teachers, 9 of 24 primary raters scored 1 teacher, 9 scored 2 teachers, 2 scored 3 teachers, 3 scored 4 teachers, and 1 primary rater scored 5 teachers; among the 32 external raters who fully scored this sample of 50 teachers, 20 external raters scored 1 teacher, 7 scored 2 teachers, 4 scored 3 teachers and 1 scored 4 teachers.

[b] Within this fully-rated cycle 2 sample (33 teachers), 6 primary raters scored a single teacher, 4 scored 2 teachers, 2 raters scored 3 teachers, 2 scored 4 teachers, and 1 primary rater scored 5 teachers; among external raters, 9 scored 1 teacher, 6 scored 2, and 4 scored 3 teachers.

[c] These 16 teachers were scored by 9 different primary raters and 10 separate external raters, the majority of whom (5 of 9 and 6 of 10) scored 1 teacher.

[d] These 27 teachers were assessed by 12 different primary raters and 16 separate external raters. Among these 12 primary raters, 5 scored a single teacher, 3 scored two teachers, 1 scored three teachers, 2 scored four teachers and 1 scored five teachers; among the 16 external raters, 9 scored a single teacher, 4 scored two teachers, 2 scored three teachers and 1 scored four teachers.

Teachers who received a more comprehensive set of scores across cycles and raters do not represent the full population of district teachers — after all, they selected into[25] the pilot process, completed it with their site supervisor during a year when many other teachers dropped out, and also had observers who tended to enter scores separately. In addition, interviews with administrators indicated that participating pilot teachers tended to be relatively experienced, hard working, and high performing. However, there is no *a priori* reason to suspect that the DFT-based observation measure and process should measure these teachers' performance *more reliably* than for other groups of teachers.[26] Put another way, it's assumed that such a system should be able to repeatedly assess teachers' practice against the DFT in an accurate way, irrespective of whether the teacher is a "low" or "high" performer.

## 6 Analysis Samples & Methods

I rely on separate teacher samples to address my research questions. As shown in table B4, these analysis samples generally featured high proportions of non-White, female elementary teachers with ten or more years of experience. The mean California Academic Performance Index (API) scores across the samples' schools ranged from 767 to 793 in 2011/12, with no sample mean reaching the statewide API goal of 800. (In comparison, the district's API for the year was 744, while the overall statewide API was 788.)

[25] Approximately two thirds of the participating administrators surveyed at the end of the 2011/12 school year indicated that their teachers independently volunteered to participate in the pilot, while other administrators encouraged these teachers take part because they felt they were well-suited to the process.

[26] Issues related to the validity of inferences based on DFT-based observation ratings, such as accurate scoring across groups or correlations with student learning, are addressed in dissertation paper C.

Table B4. Descriptive statistics for analysis samples (2011/12 data)

| Research Question: Analysis | Pilot teachers | % White | % Female | % Elem. | % teaching 10+ yrs | % with masters/ doctorate | Mean school API |
|---|---|---|---|---|---|---|---|
| RQ 1: Differences between cycles (two scores from primary rater) | 88 | 43.5 | 76.5 | 71.8 | 78.8 | 40.0 | 781 |
| RQ 1: Differences between raters/inter-rater reliability (cycle 1) | 50 | 38.8 | 73.5 | 73.5 | 81.6 | 34.7 | 767 |
| RQ 1: Differences between raters/inter-rater reliability (cycle 2) | 33 | 48.5 | 84.9 | 84.9 | 72.7 | 42.4 | 787 |
| RQ 2&3: G study: Decomposing variance across all 19 FE (both raters in both cycles) & estimating reliabilities for different numbers of rater groups/observed lessons | 16 | 50.0 | 68.8 | 75.0 | 75.0 | 37.5 | 793 |
| RQ 2&3: G study: Decomposing variance across the 10 Standard 3 FE (both raters in both cycles) & estimating reliabilities for different numbers of rater groups/observed lessons | 27 | 44.4 | 77.8 | 81.5 | 81.5 | 33.3 | 785 |
| ACROSS DISTRICT[1] | 371[2] | 41.1 | 68.4 | 65.1[3] | NA[4] | 41.4 | 744[5] |

[1] All district data besides the pilot sample size (N=371) were retrieved May 2013 from the California Department of Education (CDE) Dataquest system, online at http://dq.cde.ca.gov/dataquest/.

[2] This figure represents the number of teachers who were rated on at least one focus element by either observer in either cycle. An additional 36 teachers started the pilot process by filling out a self-assessment, but were never observed, and the district believes that at least some of these teachers, along with another 19, may have been observed during the pilot year, but ratings were never entered for them by an observer. Overall, 562 teachers received pilot training in fall 2011.

[3] This figure was derived from the total number of public elementary schools, public intermediate/middle schools, and public high schools in the district, according to CDE data for 2011/12 (from http://dq.cde.ca.gov/dataquest/).

[4] CDE data for 2011/12 indicate that the average experience level among district teachers in 2011/12 was 13.8 years.

[5] This figure represents the district's 2011/12 API, not the average API of schools in the district.

A conventional analytic approach to assessing the reliability of ratings within implemented (i.e., existing or pilot) observational systems involves examining inter-rater reliability to judge the extent to which the scores provided by different raters using the same observation instrument are free from measurement error (see, for example, Heneman & Milanowski [2003] about Cincinnati's TES or Sartain, Stoelinga & Brown [2009] about Chicago's pilot effort). Here, my research question 1 is first addressed by examining the distributions of scores assigned to the different samples of teachers with complete focus element ratings in both cycles (table B3, top row c),[27] followed by an assessment of rater agreement.

[27] For the samples of teachers rated by both observers (50 in cycle 1 and 33 in cycle 2) I examined rater agreement separately for the full set of 19 focus elements as well as for only the 10 focus elements under *Standard 3: Instruction*, in an attempt to discern whether observers may have been more consistent in assessing elements of performance in the classroom only (standard 3) or for the full suite of teacher responsibilities that define effective teaching in the district (at least during the pilot year).

Rater agreement was assessed separately during cycles 1 and 2 via Cohen's (1960, 1968) kappa coefficient for categorical data. The minimal number of teachers scored by each rater during the pilot year (only one to three in most cases, as shown in table B1) necessitated estimating inter-rater reliability at the observer group level (i.e., primary rater versus second rater) rather than at the individual rater level. Cohen's kappa coefficient ranges from 0 to 1 and is a function of the ratio of agreements to disagreements in relation to expected frequencies:

$\kappa = [\Pr(o) - \Pr(e)] / 1 - \Pr(e)$.

Where Pr(*o*) is the relative observed agreement among raters, and Pr(*e*) is the hypothetical probability of chance agreement (the observed data is used to calculate the probabilities of each observer randomly selecting each category).[28]

In reality, however, rater agreement is only one aspect of reliability. Teachers' behavior and performance can vary greatly across lessons (see Medley & Mitzel, 1963, for some of the earliest evidence of this), and conventional inter-rater reliability coefficients fail to estimate any interactions between raters, teachers, and lessons — for example, whether some raters may assign harsher grades to certain groups of teachers (Hill, Charalambous & Kraft, 2012). To address another dimension of score reliability, I rely on the tools of G theory (Brennan, 2001; Shavelson & Webb, 1991) to address research question 2 and decompose the variation in participating pilot teachers' total (sum) scores[29] to simultaneously examine different sources of variance and reveal the relative influence of teachers, rater groups (site supervisor versus second reviewer), observation occasions (lessons), and all associated interactions.[30] The tools of G theory are relevant here because "when one is concerned about the consistency of examinee rank-ordering across two testing occasions and across different raters... G theory provides reliability estimates accounting for both sources of (variation) simultaneously" (Sawaki, 2010: 533).

As shown in tables B4 and B5, my generalizability analysis examines the performance of two samples of participating pilot teachers: (1) the 16 teachers rated on all 19 focus elements by both

[28] Although there are no published statistical standards for "acceptable" levels of rater agreement, it is generally accepted in measurement circles that 0.7 represents good agreement, while above 0.8 represents very good agreement (e.g., Bell et al., 2012; DeVellis, 1991; Pallant, 2007). However, such thresholds may be unrealistic in practice, particularly in pilot years. In examinations of rater agreements in two early implementations of FFT-based observation systems, observed agreements were below .8 in Cincinnati (Heneman & Milanowski, 2003; Milanowski, 2011) and below .6 in the Chicago pilot's first year (Sartain et al., 2009).

[29] Generalizability (G) theory can be viewed as a conjunction of classical test theory (particularly the Spearman-Brown approach relating reliability to test length) and analysis of variance (ANOVA). Unlike a classical reliability analysis, which tends to consider one source of variation at a time — for example, test-retest reliability examines day-to-day variation, but not variation due to item sampling, while internal consistency counts variation due to item sampling but not day-to-day variation — G theory uses analysis of variance to partition multiple sources of variation into different sources in a single analysis. First, repeated-measures ANOVA quantifies the amount of variance associated with each facet/factor (and any interactions), and these sources of variance are used to determine which of the facets/factors (or interactions) contribute the most to measurement error. Using the resulting generalizability (reliability) coefficients, decisions then can be made regarding the manipulation or control of the sources of variance.

[30] The G theory framework considers all possible combinations of teachers, lessons (occasions) and observers to be random selections from the "universe" of possible admissible measurements (Shavelson, Webb & Rowley, 1989; Newton, 2010).

observers in both observation cycles, and (2) the 27 teachers (i.e., the 16 previous and 11 others) accordingly rated on the ten focus elements from Standard 3: Instruction.[31] I focus on these samples for two reasons. First, assessing the impact of both the number of raters and the number of observation occasions (lessons) — perhaps the two most policy-relevant facets of observation systems — required examining teachers whose scores spanned both dimensions, which was only the case for a very limited number of pilot teachers.[32] Also, recent studies of the implementations of Danielson FFT-based observation systems in Cincinnati (Kane et al., 2011) and Chicago (Sartain et al., 2011) have tended to focus only on those components/elements that address classroom practices; hence my examination of the subset of DFT Standard 3: Instruction focus elements.[33] Finally, small-sample studies of how different sources of variation may affect the results of new teacher observation systems are not uncommon. Examples can be found in Hill, Charalambous and Kraft (2012) and Martinez, Borko and Stecher (2011), which reported on G studies of measures of teacher quality that examined only 8 and 11 teachers, respectively.[34]

[31] The 10 focus elements (FE) from Standard 3: Instruction tended to be among the most commonly rated elements across raters and cycles; in particular these 10 were among the top 14 with ratings by both observers in both cycles. Examining only the 10 most commonly rated focus elements across raters and cycles would broaden the sample by only two additional teachers. Instead, I focus on Standard 3 because it is a conceptually independent domain that in itself characterizes key aspects of teachers' pedagogical performance.

[32] My G study sample of pilot teachers (n=27) tended to work in schools with proportions of English learners (32%) and disadvantaged students (79%) slightly higher than the district-wide averages (27% and 77%, respectively), and seven of the sample teachers' twelve schools had 2012 API scores above the district API of 744. [Source: CDE Dataquest system data (http://dq.cde.ca.gov/dataquest/), retrieved May 2013.]

[33] I do not also include focus element ratings for DFT Standard 2: Classroom Environment here because this domain was only assessed during the pilot year by two focus elements (although the standard contains 14 in total), and according to Herman, Heritage and Goldschmidt (2011) a minimum of five "items" is generally necessary to get a reliable score for any explicit test target, such as understanding of a particular concept or the ability to apply a specific skill. Since Standards 2 and 3 are considered conceptually independent domains on the DFT, each would generally necessitate its own sufficient item pool for reliability analysis.

[34] There is no simple answer to how many teachers are needed to carry out a G study of observation scores because the sampling error for any particular variance component depends on a number of factors. One cannot easily compute a "minimum sample size" as might be done in a power analysis for an impact study. As Smith (1981) explained, "sampling errors of variance components are for the most part dependent on four specific design characteristics: the sample size used, including both the number of levels of the facet of interest and the number of observations on each of those levels; the complexity of the expressions for the expected mean squares and accompanying computation of variance component estimates; the design configuration, e.g. nesting versus crossing; and the relative magnitudes of the population values of variance components. The relationship between these factors are by no means simple, but in general the larger the sample size and the less complex the design, the more stable the variance component estimates resulting from the design are likely to be" (p. 147). As noted, I believe the two facets examined here (observation occasions and observer groups) are both of primary significance for policymakers.

Table B5. Numbers of teachers rated by observer groups in two G study samples (n=16, n=27)

| 19 focus elements, both cycles | Total | Rated 1 teacher | Rated 2 teachers | Rated 3 teachers | Rated 4 teachers | Rated 5 teachers |
|---|---|---|---|---|---|---|
| Teachers | 16 | - | - | - | - | - |
| Primary raters | 9 | 5 | 2 | 1 | 1 | 0 |
| External raters | 10 | 6 | 2 | 2 | 0 | 0 |
| **10 focus elements, both cycles** | **Total** | **Rated 1 teacher** | **Rated 2 teachers** | **Rated 3 teachers** | **Rated 4 teachers** | **Rated 5 teachers** |
| Teachers | 27 | - | - | - | - | - |
| Primary raters | 12 | 5 | 3 | 1 | 2 | 1 |
| External raters | 16 | 9 | 4 | 2 | 1 | 0 |

As is shown in table B5, a large proportion of pilot observers only scored one teacher during the pilot year. As a result, I assess raters' impact at the group rather than the individual level, i.e., raters were grouped into one of two classifications for analysis — either primary (supervising administrator) or second (typically off-site) rater. These groups are qualitatively different because, although trained similarly and possessing similar backgrounds (table B1), the second raters generally lack a daily working (if not personal) relationship with the teacher. Given recent evidence of leniency bias towards teachers in the observation ratings from supervising administrators at their site (Sartain et al., 2011; Ho & Kane, 2013), this source of variation is a meaningful one to examine.

In G study terms, my analysis for research question 2 features a teacher-by-rater type-by-occasion (lesson) design, with teachers' total scores (summed across either all 19 pilot focus elements or the ten Standard 3 focus elements) crossed with rater type (two levels), crossed with observation occasion (two levels), with only the latter considered a random facet (or random factor in ANOVA parlance).[35] The total variability in teachers' total (sum) scores is decomposed into these two policy-relevant facets (rater groups and observation occasions),[36] their interactions, and residual error (the portion of the total score not explained by the other effects). The relative magnitudes of the variance components provide information about the particular sources of variation[37] and allow for the estimation of generalizability (reliability) coefficients, referred to as either the G coefficient (for relative or rank-order decisions) or the phi coefficient

[35] Although all teachers were not observed on the same day on each of the two occasions, here "occasions" is defined more in terms of the number of lessons observed on different days (i.e., the same class over multiple occasions).

[36] Since the object of measurement (teachers) is not a source of error, it is not considered a facet. Nor is the number of DFT elements, components or standards, despite the fact that, as noted by Hill, Charalambous and Kraft (2012), the number of individual rating prompts on a rubric essentially signifies the cognitive load for raters and is a key design issue. The number of such prompts indeed varies across many of the observation frameworks and rubrics in use today, but in their recent paper Hill and colleagues (2012) reported finding "no studies of how the quantity of items on an instrument might affect raters' performance and consequently the characteristics of the resulting teacher scores" (p. 58). Moreover, discussions with district officials indicated that their set of focus elements was established through careful deliberation and as such are not likely to change drastically (district officials, personal communication, 2012). Thus this issue isn't explored in this analysis of pilot-year ratings.

[37] Estimates of the variance components are obtained from analysis of variance by setting the expected mean squares for each variance component equal to the observed mean squares and solving the set of simultaneous equations (not displayed).

(for absolute, criterion-referenced decisions).[38] These coefficients provide a means, in turn, to address my third research question, which examines how changes to the design of the measurement system — here the number of observation occasions (lessons) and/or observer groups — might achieve certain levels of reliability for relative or absolute decisions.[39]

# 7 Results

*RQ1. How consistent across observation cycles and raters were the DFT-based scores assigned to pilot teachers?*

Table B6 displays the percentage distribution of total focus element scores awarded to the 88 participating pilot teachers who were scored by their primary rater in both observation cycles, while table B7 displays the percentage distribution of focus element scores for the separate samples of teachers who were fully scored by both rater groups during cycle 1 (50 teachers) and cycle 2 (33 teachers) and across both cycles (16 teachers). The means and distributions displayed are calculated not across individual teachers but across all ratings assigned to those teachers.

Table B6. Percent distribution of total focus element scores awarded to the sample of 88 participating pilot teachers by the primary rater who scored them in both observation cycles

| All Standards: 88 teachers x 19 FE | Mean (SD) | %Ineff. | %Developing | %Effective | %Highly Eff. |
|---|---|---|---|---|---|
| Cycle 1 | 2.74 (0.42) | 1.1 | 32.7 | 58.0 | 8.3 |
| Cycle 2 | 2.87 (0.46) | 0.4 | 25.5 | 60.3 | 13.8 |
| **Standard 3: Instruction 88 teachers x 10 FE** | **Mean (SD)** | **%Ineff.** | **%Developing** | **%Effective** | **%Highly Eff.** |
| Cycle 1 | 2.65 (0.47) | 1.9 | 38.3 | 52.6 | 7.2 |
| Cycle 2 | 2.81 (0.47) | 0.7 | 29.0 | 58.9 | 11.5 |

[38] The relative versus absolute distinction relates to the components of the error terms. All variance components except for the object of measurement (teachers) enter the error term for absolute decisions, while all variance components affecting relative standing (i.e., including teacher effects) enter the error term for relative decisions (Newton, 2010).

[39] For relative decisions in a crossed design, the effect of observer group and cycle (lesson) are constant for all teachers and so don't influence the rank ordering; thus the three variance components associated with only observer group and cycle are not included in the error variance for relative decisions (Shavelson, Webb & Burstein, 1986).

Table B7. Percent distribution of total focus element scores awarded to common teachers by primary and secondary raters in cycle 1 (50 teachers), cycle 2 (33 teachers) and across both cycles (16 teachers)

| **Cycle 1 (50 teachers)** | | | | | |
|---|---|---|---|---|---|
| All standards: 50 teachers x 19 FE | Mean (SD) | %Ineff. | %Developing | %Effective | %Highly Eff. |
| Primary observer [24] | 2.69 (0.36) | 1.2 | 33.8 | 60.7 | 4.3 |
| Second observer [32] | 2.63 (0.35) | 1.1 | 36.8 | 59.1 | 3.1 |
| Standard 3: Instruction 50 teachers x 10 FE | Mean (SD) | %Ineff. | %Developing | %Effective | %Highly Eff. |
| Primary observer [24] | 2.59 (0.40) | 2.0 | 40.0 | 55.0 | 3.0 |
| Second observer [32] | 2.55 (0.40) | 2.0 | 42.8 | 53.0 | 2.2 |
| **Cycle 2 (33 teachers)** | | | | | |
| All standards: 33 teachers x 19 FE | Mean (SD) | %Ineff. | %Developing | %Effective | %Highly Eff. |
| Primary observer [15] | 2.95 (0.44) | 0.8 | 19.8 | 63.6 | 15.8 |
| Second observer [19] | 2.92 (0.45) | 0.6 | 22.5 | 60.6 | 16.3 |
| Standard 3: Instruction 33 teachers x 10 FE | Mean (SD) | %Ineff. | %Developing | %Effective | %Highly Eff. |
| Primary observer [15] | 2.87 (0.48) | 1.2 | 24.2 | 60.6 | 13.9 |
| Second observer [19] | 2.86 (0.51) | 1.2 | 25.8 | 57.9 | 15.2 |
| **Both Cycles (16 teachers)** | | | | | |
| Cycle 1: 16 teachers x 19 FE | Mean (SD) | %Ineff. | %Developing | %Effective | %Highly Eff. |
| Primary observer [9] | 2.69 (0.42) | 1.3 | 34.9 | 57.6 | 6.3 |
| Second observer [10] | 2.63 (0.36) | 1.3 | 36.8 | 59.2 | 2.6 |
| Cycle 2: 16 teachers x 19 FE | Mean (SD) | %Ineff. | %Developing | %Effective | %Highly Eff. |
| Primary observer [9] | 2.94 (0.51) | 1.3 | 21.7 | 58.6 | 18.4 |
| Second observer [10] | 2.89 (0.48) | 1.0 | 24.3 | 58.9 | 15.8 |

The general expectation in designing observational rubrics is that performance levels will accurately and adequately capture teacher performance and that scores will span all rubric levels (McREL, 2012). If raters only ever used one or two points on a scale that had four levels (as the DFT does), then one might question the degree to which the descriptions for the levels were appropriate (Bell et al., 2012). As shown in tables B6 and B7, the DFT scores awarded during the pilot year generally spanned all levels, although few Ineffective ratings were assigned and pilot teachers' scores tended to cluster at the higher end of the performance scale (Effective or Highly Effective) than at the lower end (Ineffective or Developing).[40] Some balance was evident, though. Not everyone got high scores. Although over half of the scores given were 3's (Effective), the next most common rating was 2 (Developing), which tended to comprise 20 to 40 percent of scores across raters and cycles.

The pool of instruction-focused DFT elements (Standard 3) yielded higher proportions of Developing (Level 2) scores, slightly lower mean scores (-2–4 percent), and slightly higher

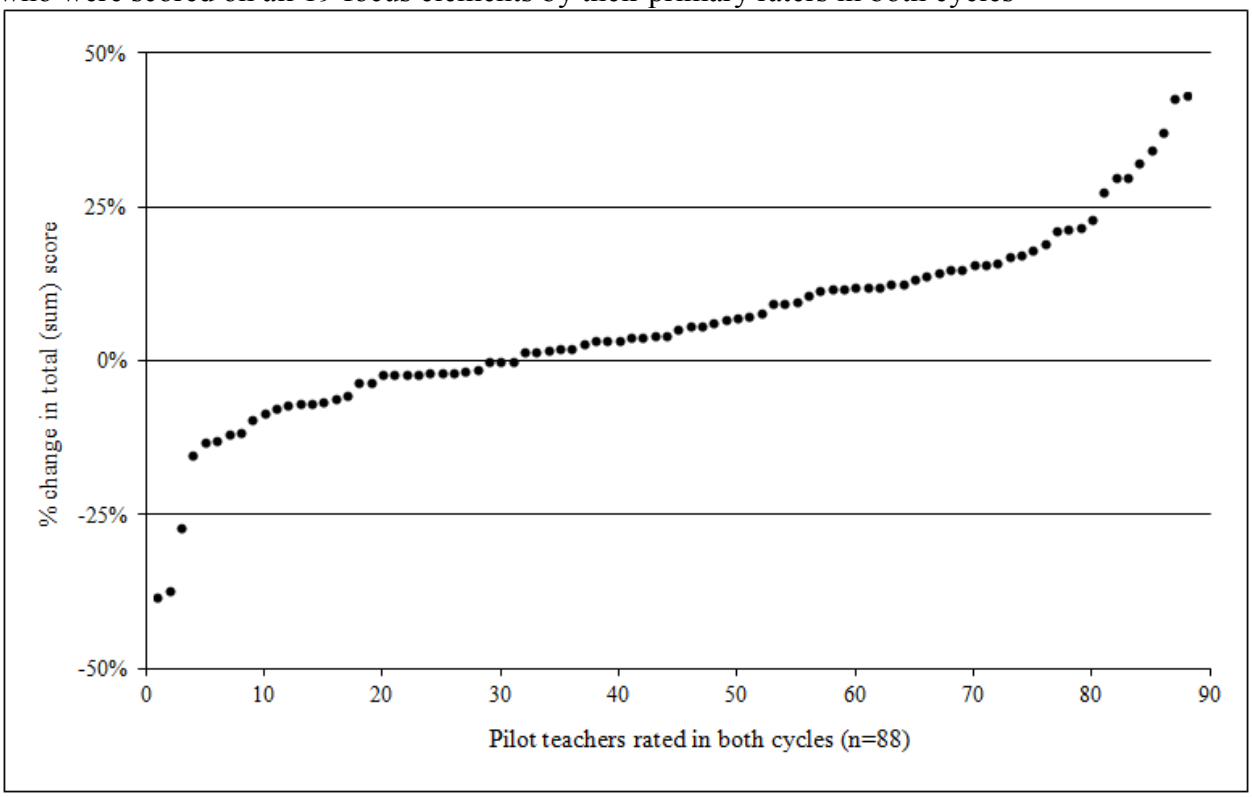
[40] This evidence does not necessarily undermine the DFT scoring design, because it might be reasonable for actual teaching practice to be clustered around particular score points for this sample of teachers (as noted, discussions with site administrators suggested that these teachers tended to be experienced, hard working, and high performing), and the distribution of "true" teacher performance is not necessarily definable. However, if the scoring design is valid then one would expect to see additional converging evidence over time (e.g., from videos coded by expert/master observers) indicating that such skewed scores reflect actual instructional quality.

standard deviations (+9–14 percent) across cycles and rater groups. The latter result indicates that raters were more likely to differentiate among teachers on these classroom-based instructional practices and use extreme values, particularly the lowest category. Of the 30 total Ineffective (DFT Level 1) scores awarded by raters across the pool of commonly-rated teachers in cycles 1 and 2, 28 (93 percent) were awarded for elements under Standard 3: Instruction. Four instructional focus elements in particular seemed to prove difficult for pilot teachers: 3b1 Quality and Purpose of Questions; 3b2 Discussion Techniques; 3d1 Assessment Criteria; and 3d4 Student Self-Assessment and Monitoring of Progress. Together, these four focus elements yielded more Ineffective or Developing scores than Effective or Highly Effective scores from raters in cycle 1, and across the two cycles 22 of the 30 (73 percent) Ineffective (DFT Level 1) scores were awarded for performance on these four elements.

Differences Between Observation Cycles 1 and 2

Across rater groups, teachers tended to be scored higher during the second observation cycle (March to May) than during the first cycle (October to February), with higher percentages of Effective or Highly Effective scores and slighter higher overall mean scores. The overall mean scores across the 88 pilot teachers who were scored on all 19 focus elements by their primary rater in both cycles increased 0.32 standard deviations from cycle 1 to cycle 2, although not all teachers improved. Fifty-seven of these 88 participating pilot teachers (65 percent) saw an increase in their total (sum) scores from their primary raters, from simply gaining one level on one focus element to an overall improvement of over 40 percent (figure B1). Conversely, 28 of these 88 teachers (32 percent) received lower total (sum) scores from their primary raters across the 19 focus elements in cycle 2 than they did in cycle 1, with a fairly parallel range of differences. Figure B1 displays the percentage change in the total (sum) scores for each of the 88 pilot teachers with 19 focus element scores from their primary rater in both cycles. (Each point represents a teacher.) Between cycles 1 and 2 the mean change in these teachers' total (sum) scores from primary observers was +6.0 percent, and the median change was +4.9 percent.

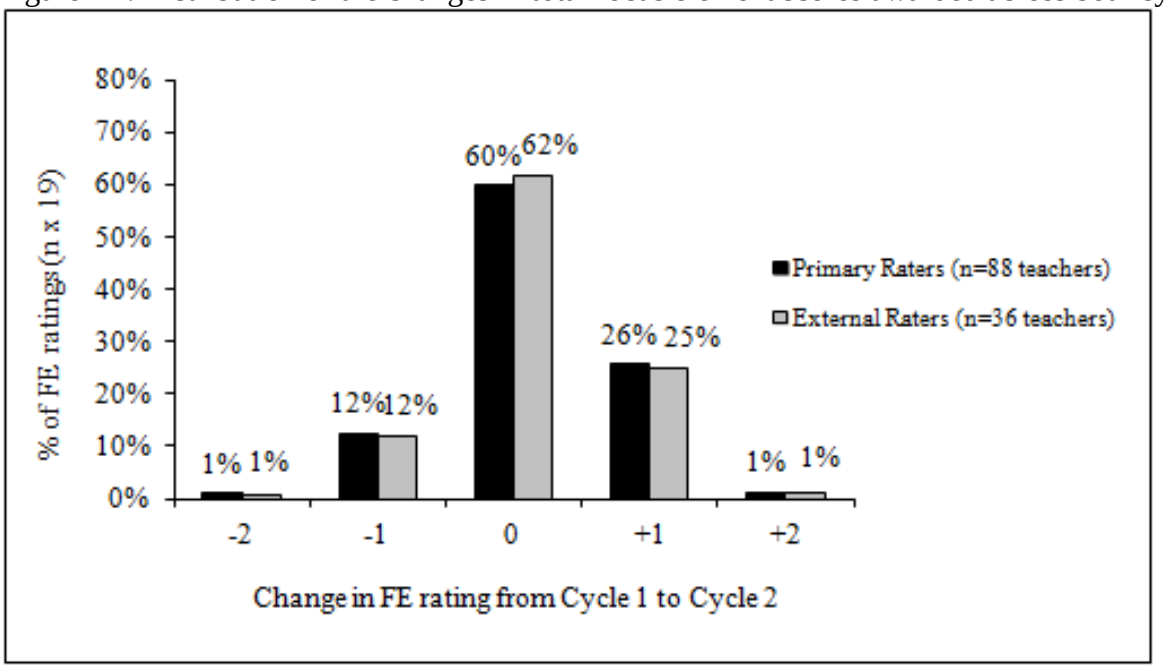
Figure B1. Distribution of changes in total (sum) scores of the individual pilot teachers (n=88) who were scored on all 19 focus elements by their primary raters in both cycles




If, in each cycle, one were to classify these 88 pilot teachers by their average score (with performance cutoffs at 1.0, 2.0, and 3.0, for example), then 21 of the 88 teachers (24 percent) saw their classifications increase one performance level from cycle 1 to cycle 2, while seven teachers (8 percent) saw their classifications drop one level. Twenty-five of these 28 classification shifts pushed the teacher below or above 3.0, between the Developing and Effective levels, hypothetically. (The study district has not yet established any "cut score" for satisfactory/unsatisfactory performance on the DFT.)

Figure B2 displays the distribution of the changes among the individual focus element scores awarded from cycle 1 to cycle 2, for those pilot teachers who were scored on all 19 focus elements on both occasions. Primary raters awarded this sample of 88 teachers a total of 1672 FE scores in each of the two cycles, while external raters scored only 36 teachers on all 19 focus elements in both cycles (for a total of 684 FE scores in each cycle). As shown, across the rater groups approximately 60 percent of FE scores remained the same in each cycle, while about 25 percent increased by one DFT level on the 1-4 scale and about 12 percent decreased by a DFT level from cycle 1 to cycle 2.

Figure B2. Distribution of the changes in total focus element scores awarded across both cycles




Note: The distributions in this graph are comprised of 1672 scores (cycle differences) from primary raters and 684 scores (cycle differences) from external raters across the two cycles.

Differences Between Rater Groups

Across cycles, the mean scores from external raters tended to be slightly lower (by .08 to .19 standard deviations) than those awarded by primary raters, with external raters awarding higher proportions of Developing (Level 2) scores to commonly-rated teachers. Figure B3 plots the mean scores from the two groups of raters against one another. The top panel plots the mean scores from the 50 teachers rated on all 19 focus elements by both groups of observers in cycle 1, while the bottom panel plots the mean scores from the 33 teachers thusly rated in cycle 2.[41] Each point represents a pilot teacher — a total of 77 different teachers are displayed (as 16 teachers are in both groups) — and the red line represents the 45-degree line where the two scores would be equivalent. The average score from the primary rater (the teacher's supervising administrator) is on the vertical axis and the average score from the second (typically off-site) rater is on the horizontal axis. As shown, although many of the scores match, more of the score discrepancies lie above the 45-degree line, meaning that teachers' supervising administrators tended to provide higher mean scores than second raters.

[41] Again, towards the end of the pilot year the district asked observers to agree on a single set of focus element ratings; however, these 33 teachers had two sets of ratings entered in cycle 2.

Figure B3. Comparing mean scores (across 19 FE) from primary and external raters in cycle 1 (50 common teachers) and cycle 2 (33 common teachers)

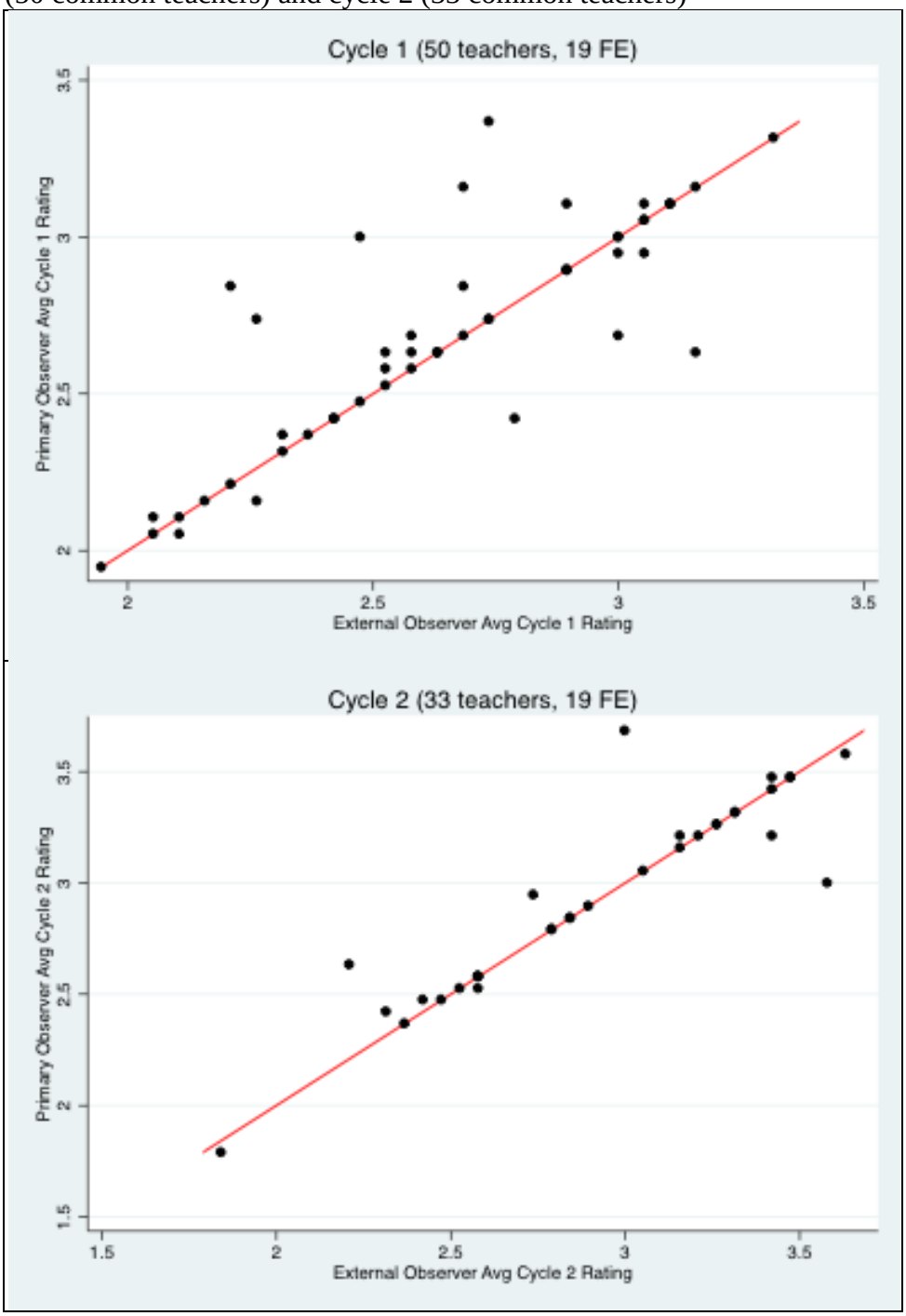


Table B8 displays the reliability (kappa) coefficients for the two rater groups scoring common pools of teachers in cycle 1 (50 teachers) and cycle 2 (33 teachers). These two different samples of teachers both include the 16 individuals scored by both rater groups in both cycles (i.e., 77 individuals are represented).

Table B8. Inter-rater reliabilities (kappa) for different rater groups in cycles 1 and 2

| Primary vs External Rater | Cycle 1 (n=50) | Cycle 2 (n=33) |
|---|---|---|
| All Standards [19 FE] | .714 | .763 |
| Standard 3: Instruction [10 FE] | .724 | .802 |

Note: The sample of 50 fully-rated cycle 1 teachers were scored by 24 different primary raters and 32 different second raters; the sample of 33 fully-rated cycle 2 teachers were scored by 15 different primary raters and 19 different second raters. The unadjusted proportions of rater agreement were .853 (across 19 FE) and .852 (across 10 FE) in cycle 1 and .871 (19 FE) and .888 (10 FE) in cycle 2.

As shown in table B8, good agreement (i.e., 0.7 or higher [DeVellis, 1991; Pallant, 2007]) was evident between primary and second raters in both cycles, with a slightly higher level of rater agreement evident among the ten elements assessing Standard 3: Instruction, suggesting that although raters generally awarded lower (and more varied) scores for these classroom-focused elements, they may have had a more aligned understanding when rating teachers' performance in these areas. (The differences between the scores from the two groups of raters were 1/3 to 1/2 a standard deviation smaller for this subset of elements, as shown in table B7.) The level of rater agreement increased from cycle 1 to cycle 2, which is not necessarily surprising given that observers had met and collaborated more by that stage of the year. When asked on end-of-year surveys to select the option that best represented how they entered formal observation scores online during cycle 2, 64 percent of responding primary and second raters[42] said they entered teachers' scores separately after formal observations, while 36 percent indicated that on average they collaborated and entered a single set of agreed-upon scores.[43] Therefore, many participating pilot teachers may have been observed and given feedback by their two observers, just not separately scored by them, and these teachers (and observers) are excluded from the analyses of rater agreement and score variation presented here.

[42] The end-of-year survey response rates among these two observer groups were 51 percent for participating pilot site administrators (64 of 125) and 45 percent for participating second observers (75 of 167).

[43] In theory, an agreed-upon rating from two observers is likely to be a more reliable and accurate (i.e., less biased or noisy) rating of teacher performance.

*RQ2. How much of the variance in pilot teachers' total DFT scores was attributable to systematic differences between teachers?*

Table B9 reports the distinct sources of variance in participating pilot teachers' total (sum) scores across either the full 19 focus elements (n=16) or only the ten focus elements from Standard 3: Instruction (n=27).

Table B9. Estimated variance components (with relative magnitude) and generalizability coefficients for total (sum) scores among two G study samples (n=16, n=27)

| | 19 FE [n=16] | | 10 FE [n=27] | |
|---|---|---|---|---|
| Source of variation | Est. variance component | % total variance | Est. variance component | % total variance |
| Teacher | 56.5 | 67.8 | 13.8 | 57.8 |
| Cycle | 11.4 | 13.7 | 3.1 | 13.1 |
| Observer group | 0.2 | 0.2 | 0 | 0 |
| Teacher x Cycle | 9.6 | 11.5 | 6.1 | 25.2 |
| Teacher x Obs group | 4.8 | 5.7 | 0.6 | 2.5 |
| Cycle x Observer group | 0 | 0 | 0 | 0 |
| Residual | 0.8 | 1.0 | 0.3 | 1.4 |
| Generalizability coefficient | | | | |
| Relative (G) | 0.884 | | 0.820 | |
| Absolute (phi) | 0.811 | | 0.752 | |

Note: The G coefficient refers to the measurement consistency of relative decisions comparing teachers against one another, while the phi coefficient refers to the consistency of absolute decisions irrespective of how other teachers score.

The first row refers to the variance attributable to teachers, which accounts for 68 percent of the total variance in teachers' total (sum) scores across all 19 focus elements, and 58 percent of the variance across the 10 Standard 3 focus elements. This suggests that systematic differences between teachers were fairly well identified in these samples, and that the 19 focus elements captured more variation attributable to teachers in the pilot year than did the ten Standard 3 focus elements. Next shown is the variance attributable to differences between observation occasions (cycles) — two observed lessons delivered by the same teacher, the first between October and February and the second between March and May. As shown, together the variance components associated with cycle overall and the teacher-cycle interaction (i.e., particular teachers systematically performed better or worse in particular cycles) combined to account for over 25 percent of the variance in teachers' total scores across the 19 focus elements. Examining teachers' total scores based on just the ten Standard 3 focus elements, the proportion of variance attributable to the cycle alone was similar to that in the 19-element sample (13.1 versus 13.7 percent), but the variance attributable to the teacher-cycle interaction was much higher (25.2 versus 11.5 percent), suggesting that the relative standing of certain teachers on these elements tended to change more from one cycle to another. Teachers did tend to improve between cycles, as noted, but such growth wasn't universal (see figures B1 and B2). Ultimately, growth and development during the teacher's EGDC year is a core goal of this reform, and it seems some participating pilot teachers made progress in 2011/12.

The direct and interaction effects associated with observer group accounted for very little of the overall variation in teachers' total scores (only about six percent in the 19 FE sample and zero in the Standard 3 [10 FE] sample), suggesting little systematic variation between scores from the two rater groups, especially among the ten classroom-focused elements in Standard 3. Ratings

were well calibrated at each cycle: the rater group-by-cycle interaction was zero. This is not necessarily surprising, however, as district guidance and end-of-year survey results suggested that many pilot raters collaborated together before entering scores. In such situations we would expect this variance component to be lower than if the two raters had been completely independent from one another, as was the case, for example, in the videotaped lessons in the MET project studies (and perhaps an unrealistic scenario in practice). The policy implications of this are discussed further in Section 8 of this paper.

*RQ3. Using these pilot year results, how might a different number of observers or observation cycles influence the reliability of teachers' total DFT scores?*

Using the G-study results from these particular pilot-year observation data — in which two raters scored two lessons from a small sample of teachers — my subsequent analysis computed results for different combinations of observation cycles and rater groups in order to forecast various estimates of reliability (table B10). Since the variability in total scores associated with cycles was far higher than that associated with rater groups (table B9), the pilot data suggest that observing additional lessons per teacher would increase estimated reliability far more than having additional rater groups score the same teachers (and, in parallel, that reducing the number of rater groups would decrease estimated reliability less than reducing the number of observations).

Table B10. Estimated reliability coefficients for different numbers of rater groups and observed lessons (cycles), based on G study results from total (sum) scores among the two G-study samples (n=16, n=27)

| Observed lessons | 1 | 1 | 2 | 1 | 1 | 3 | 4 | 2* | 2 | 3 | 4 | 3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rater groups | 1 | 2 | 1 | 3 | 4 | 1 | 1 | 2* | 3 | 2 | 2 | 3 |
| 16 teachers (19 FE) | | | | | | | | | | | | |
| Relative (G) | .78 | .82 | .85 | .83 | .84 | .87 | .88 | .88 | .89 | .90 | .92 | .92 |
| Absolute (phi) | .67 | .70 | .75 | .71 | .72 | .82 | .83 | .81 | .82 | .85 | .87 | .86 |
| 27 teachers (10 FE) | | | | | | | | | | | | |
| Relative (G) | .66 | .68 | .78 | .68 | .69 | .83 | .85 | .82 | .82 | .85 | .88 | .86 |
| Absolute (phi) | .57 | .59 | .72 | .59 | .60 | .78 | .82 | .75 | .75 | .80 | .84 | .81 |

Note: Figures are based on a fully crossed Teacher x Rater Group x Lesson (Cycle) x Total Score design, where only lesson (cycle) is considered a random facet. The G coefficient refers to the measurement consistency of relative decisions comparing teachers against one another, while the phi coefficient refers to the consistency of absolute decisions irrespective of other teachers' scores.

* Represents the pilot-year observation design (teachers scored on two lessons by two rater groups).

The estimated reliability coefficients for participating pilot teachers' total (sum) scores under the pilot-year observation design (two lessons observed by two rater groups) ranged from 0.81 to 0.88, indicating that, at least for these small samples of pilot teachers, the DFT-based observations had promising measurement properties. These results also suggest that, given the way the observation system was implemented in the pilot year (and given the way the resulting data were analyzed here), the number of observations, rather than rater groups, represents a more efficient way to increase reliability and measurement consistency. In these data, the percentage gain in reliability from having a single rater conduct an additional observation of a teacher is about twice as large as the gain from having an additional rater observe the same lesson. This is

due to the fact that the cycle-to-cycle variance was far greater than the corresponding variance for raters (which was virtually nonexistent during the pilot year). Although this result runs counter to recent MET study findings, which concluded that adding raters was the best way to increase reliability in observation scores (Ho & Kane, 2013), the MET research involved outside observers scoring videotaped lessons rather than conducting actual visits to school sites. Multiple independent raters may indeed represent a worthy safeguard against personal biases, but absent widespread videotaping (which is generally unrealistic in practice) they tend to impose logistical challenges, and this burden was clearly evident in the participant feedback received during the pilot year (Strunk et al., 2013b).

Districts working to implement these types of systems must balance competing priorities and make tradeoffs, for example weighing the technical precision afforded by "adding another set of eyeballs" (Ho & Kane, 2013: 22) against ensuring a manageable and fruitful professional feedback process in practice. In this context, because very little of the variation in participating pilot teachers' total (sum) scores was attributable to raters, a policy shift from, for example, two observers to a single observer (still conducting at least two observations) could reduce scheduling burdens while also maintaining a high level of estimated reliability in teacher rank-ordering. As long as at least two observations are conducted, these particular data from the pilot year suggest that the estimated reliability of relative distinctions between teachers remains above the 0.80 benchmark in the one rater, two-observation scenario displayed in table B10.

## 8 Limitations of this research

Pilot phases of major reforms are not ideal contexts for measurement studies. Here the tools under study were still being revised and fine tuned during the pilot year; observers were still learning the tools and teachers and administrators were just becoming familiar with the processes and measures. These results are also based on a small sample of motivated, compensated volunteers who carried out the observation process in a particular way during a pilot year that featured a high level of attrition. These participating teachers and administrators may differ from non-participants in other key ways beyond this pilot-specific motivation and context. Given this situation, it's possible that the reliability of DFT ratings during the pilot year may not reflect what will be found in an eventual full-scale roll out.

## 9 Discussion

This paper examined the reliability of results from the pilot implementation of a modified version of the Danielson Framework for Teaching with volunteering teachers and administrators in a large California school district. Participating pilot teachers' observation scores were analyzed in various ways — first examining the distributions of scores by observation cycle and rater group (site supervisor versus second rater), then conducting a traditional analysis of rater agreement, and finally, utilizing a G theory framework to both estimate the extent to which raters and observation occasions (lessons) impacted scores during the pilot year as well as forecast how

different combinations of rater groups and observed lessons might influence the reliability of those scores (based purely on results from pilot-year implementation).

Although adjustments were made throughout the pilot year and only a small sample of teachers received a complete set of focus element scores from both raters across both cycles, results from these teachers indicated that they tended to be scored higher during the second cycle (although such improvement wasn't universal), and across cycles the scores from second raters (who typically did not work at the school site) tended to be slightly lower than those awarded by the teachers' supervising site administrator. Participating teachers also tended to receive lower scores for the ten focus elements assessing DFT Standard 3: Instruction, with elements 3b1, 3b2, 3d1 and 3d4 particularly characterized by lower scores (especially in the first observation cycle).

Good agreement was evident between the primary and second raters who scored common teachers, with a slightly higher level of rater agreement evident for the ten Standard 3 focus elements (perhaps suggesting more aligned understanding) as well as during the second observation cycle (when raters had likely met and collaborated more). Although the district initially afforded observers discretion in how they input teacher scores — in some cases only primary observers input scores, other observer teams conferred before entering a single set of scores, and in other instances each observer input his or her ratings independently — late in the pilot year the district guided observers to collaborate and enter a single-set of agreed-upon scores. This paper, however, tended to focus on the results for teachers who had scores from two raters, and the findings presented her therefore should not be thought of a representative of all pilot participants.

Generalizability analyses showed that approximately two-thirds of the variation in participating pilot teachers' total (sum) DFT scores was attributable to systematic differences among teachers, while the variability associated with the observation cycle (approximately 25 percent) was larger than that associated with rater group (approximately 6 percent). These results were then used to forecast reliability coefficients based on different combinations of rater groups and observed lessons (cycles), and indicated that, based solely on pilot implementation and results from this particular analysis sample, varying the number of observations influenced reliability estimates far more than varying the number of observers. In particular, as long as at least two observations are conducted, these particular data from this district's pilot year suggest that the reliability of relative distinctions between teachers could remain above the traditional 0.80 threshold with a single rater. From a policy perspective, such a finding can help the district balance its competing priorities moving forward with its standards-based, multiple-measure teacher evaluation system. Coordinating second observers during the pilot year threatened the viability of the DFT system (Strunk et al., 2013b), and videotaped scoring (as in the MET research) tends to be unrealistic in practice.

Ultimately, observations of teacher practice tend to be resource-intensive processes, and the study district, like others across the country, is under a tight policy timeline for implementation of its new standards-based multiple measure teacher evaluation reform. These results also have a degree of external validity because the observations were conducted as they would be in many other large urban districts — by school and district administrators during actual (not video recorded) lessons and with participating administrators tending to know the evaluated teachers

and often each other. Findings from this large California district, an organization with a complex history and vast array of diverse constituents, have important implications for large urban districts nationwide considering similar reforms.

This paper should caution those now sprinting to develop new systems of teacher observation to think carefully about not only the instruments and processes involved but also the key contextual aspects of implementation. Not only must practitioners understand the technical tradeoffs involved with particular design decisions, but researchers cannot ignore the pressing time and resource constraints involved with these types of performance measures.

## Paper C: Exploring the validity of teacher observation ratings during pilot implementation in a large California school district

### 1 Introduction

Many states and school districts have recently instituted revamped teacher evaluation policies in response to federal initiatives like the Race to the Top and Teacher Incentive Fund grant programs and a changing political climate favoring holding teachers accountable for the performance of their students (Baker et al., 2013). Nearly two-thirds of U.S. states have made changes to their teacher-evaluation policies since 2009, and today 25 states require at least an annual evaluation of teachers (Jerald, 2012).

Education scholars have studied (and carried out) teacher assessment and evaluation for years. Yet today, acknowledging the multidimensional nature of the work, teacher assessment researchers within education tend to resist the application of uniform techniques or routines. In their work teachers must make use of diverse strategies "activated by sophisticated judgments grounded in disciplined experimentation, insightful interpretation of (often ambiguous) events, and continuous reflection" (Darling-Hammond & Snyder, 2000: 524). Authentically assessing such efforts in a reliable way involves regularly sampling teachers' knowledge and skills as they are applied in context and tracking multiple sources of evidence over time (see, for example, Darling-Hammond et al, 2012; Kane & Staiger, 2012). And many of today's policy overhauls indeed mandate the incorporation of multiple measures into teacher performance evaluations, and a number of local education agencies around the country are leading the way in implementation (Banchero, 2012; Doyle & Han, 2012).

These new systems tend to require multiple measures of performance, commonly including classroom observations, estimates of teachers' contributions to their students' test outcomes, and stakeholder surveys that measure parent and/or student beliefs about teacher quality. However, if results from these types of new measures are to be used to inform large-scale, high-stakes personnel decisions linked to compensation or retention, it becomes necessary to attempt to validate these measures, applying the same assessment standards in this area as have traditionally been applied to other areas of educational assessment. Yet to date there has generally been limited research published on the reliability and validity of the current generation of observation protocols that states and districts are considering for teacher evaluation purposes (Goe, Bell & Little, 2008), with much of the evidence coming from studies of observational measures generated for academic research purposes and not from systems that are being implemented in districts for current or eventual stakes.

This paper examines the validity of teacher ratings from a new observation-based measure of teacher effectiveness piloted in a large California school district during the 2011/12 school year. As elsewhere, these ratings, though piloted without any professional stakes, have been developed to provide valid information about the effectiveness of the individual teacher's instructional performance, here defined by the new District Framework for Teaching (DFT). If this were indeed the case, then participating teachers would be likely to view their pilot results as fair and accurate and one would generally expect ratings to converge with other available indicators of

teaching performance during the pilot year, while also showing no significant association with theoretically-unrelated factors like the teacher's gender, ethnicity, or grade level taught.[44]

To test the validity of these particular claims, I use classroom observation and value-added scores as well as student survey results from participating pilot teachers to address the following three research questions:

1. To what extent did participating teachers report on end-of-year surveys that the DFT-based observations of their practice conducted during pilot implementation represented a valid measure of their effectiveness?
2. To what extent did participating pilot teachers' classroom observation-based DFT ratings correlate with theoretically-unrelated teacher-level variables, including the grade span taught by the teacher and his or her gender and ethnic background?
3. Among the participating teachers assessed on multiple teacher effectiveness measures during the pilot year, to what extent did their classroom observation-based DFT ratings correlate with contemporary measures, specifically their surveyed students' perceptions of their classroom experience and value-added measures of their effectiveness?

These questions are central to the current debate around the validity of measuring teacher performance via classroom observation, and generally align with those explored in the recent large-scale Measures of Effective Teaching (MET) research project that just concluded its work across several large school districts. However, unlike the MET project, this study was not part of a controlled academic research experiment where anonymous expert observers rated videotaped lessons far from classrooms. Instead, results arose from the actual implementation of a new (pilot) policy in schools where volunteering teachers and administrators worked together and tended to know each other. And this large district provides a particularly interesting setting in which to ask these questions, as it has made clear that it is moving forward with the implementation of its standards-based, multiple-measure teacher evaluation system, and participating pilot teachers and principals were well aware that the eventual (and likely near-term) roll out of the reform would include a high-stakes component. Studies carried out in purely research-oriented settings can provide valuable lessons to the field, but they cannot entirely speak to what will be found in real-life implementations for current or eventual stakes.

At the same time, the uniqueness of this pilot also restricts the generalizability of findings. The pilot consisted primarily of volunteers, and there was some attrition during the pilot year — approximately one-third of the teachers trained in fall 2011 never had any ratings entered online by an observer (see section 4 of this paper). In turn, the final pilot sample was comprised of a self selected group of experienced, mostly elementary teachers who administrators from case

[44] However, these are not by definition necessary and/or sufficient conditions for validation. Teachers may be misguided about the effectiveness of their own performance, other performance indicators may be invalid, or there might be true correlations between gender, ethnicity and/or grade level and effectiveness. This paper simply adopts this particular, integrated line of argument to gather initial evidence about the validity of interpreting DFT ratings as fair and accurate measures of teachers' pilot-year performance. Other assumptions related to the validity of the scoring inference, such as the appropriateness of the scale (score distributions), the consistency of scoring across raters, and the degree of score variance attributable to systematic differences among teachers, were explored in the reliability-focused paper B.

study sites tended to characterize as particularly hard working and high performing, and who as a group had on average higher and less variable three-year value-added scores in English Language Arts (ELA) and Math in 2011/12 than did the full sample of district teachers with such scores ($p<.05$). Moreover, the tools under study were still being revised and fine-tuned during the pilot year; observers were still learning the tools and teachers and administrators were just becoming familiar with the processes and measures. Given this context, one might expect that the relationships between the piloted teacher effectiveness measures may not reflect what will be found in an eventual full-scale roll out. On the one hand, the strong selection bias inherent in the final pilot sample may lead to overestimation of the relationships between the piloted effectiveness measures, while the lack of consequences and the newness of the process might lead to underestimation.

The remainder of the paper proceeds as follows: Section 2 provides some background on the pilot phase of the district's new standards-based, multiple-measure teacher evaluation system. Section 3 reviews the relatively recent body of similar work examining the validity of the type of classroom observation measure piloted in this district, in particular its relationships with student survey- and test score-based measures of teachers' classroom performance. Sections 4 and 5 describe the data, analysis samples and methods applied. Section 6 presents results, and section 7 discusses the limitations of this research. The paper concludes in section 8 by relating these results to other recent research findings, concluding with implications for the district (and others implementing similar systems).

## 2 Background

Observational systems for assessing teachers generally start with the development, adoption or adaptation of a performance rubric describing a set of dimensions or domains of teaching. Teachers are then observed and rated on those domains some specified number of times by a trained observer (or observers) on a scale that has been operationally defined by a rubric that describes performance at each of the scale points (Bell et al., 2012; Herman, Heritage & Goldschmidt, 2011). This district engaged in just such a process to develop and initially implement its District Framework for Teaching (DFT).

A 2009 study found that 99 percent of teachers in the district received a meets standards rating under the district's evaluation system, and that only 64 percent of surveyed teachers reported that evaluation provided them with information and strategies they could use to improve their practice (Teacher Project, 2009).[45] In early 2010, soon after the publication of a final report

[45] The district's evaluation system, like other in California, is based on the provisions of The Stull Act, a 1971 California law that mandates that all school districts establish a uniform system for evaluating certificated personnel (The Stull Act, 1971). In the district, the Stull evaluation is guided by a checklist — based on the California Standards for the Teaching Profession — that administrators are expected to fill out when they evaluate teachers. The form outlines five standards of professional practice on which teachers are rated, including: support for student learning, planning and designing instruction, classroom performance, developing as a professional educator, and punctuality, attendance, and recordkeeping. Each standard has a series of sub-elements that provide greater detail on the specific behaviors that are required to meet the standard. Teachers must receive one of two ratings on the overall Stull evaluation: meets standard performance or below standard performance.

from the district's teacher performance task force, the local school board directed the district to develop a new system of teacher evaluation and support to address many of the concerns and suggestions raised about the district's current systems.

The district's resulting standards-based, multiple-measure teacher evaluation system (SBMMTES) was developed during the 2010/11 school year, piloted in 2011/12, and was originally communicated to be ready for 2012/13 school year. Upon scale-up, the district's SBMMTES was intended to include multiple measures of teacher effectiveness, including: (1) classroom observations of teacher practice by a site administrator and a second (external) observer, using protocols aligned with the District Framework for Teaching (adapted from the Danielson Framework for Teaching);[46] (2) stakeholder feedback surveys of students and parents; and (3) teacher-, grade- and subject-level and school-wide value-added measures of teachers' contribution to student achievement on standardized test scores.

The district spent much of the 2010/11 school year developing its District Framework for Teaching (DFT), intended to establish a common language and set of norms for effective teaching in the district, as well as to serve as the eventual foundation for teacher performance reviews and professional development within the SBMMTES. In December 2010, the district convened an ad hoc committee of over 150 teachers, instructional representatives, community partners, and outside consultants to make adjustments to Danielson Framework for Teaching in order to draft a version that was appropriate for the district. As part of this process, the district held over 60 focus groups with stakeholders throughout the district to gather feedback about the potential measure, and reconvened the ad hoc committee in March 2011 to create a final draft of the District Framework for Teaching, which includes five standards, 19 components (2-5 per standard), and 63 elements (2-4 per component).[47] Observation rubrics, templates for the lesson design, teacher self-assessments, and individual growth plans were also developed based on the DFT to be part of the SBMMTES.[48] Also during 2010/11, the district worked with a nationally-recognized research center to generate value-added measures of teacher performance, and also partnered with local university researchers to develop and field test a set of classroom and school environment surveys for students and parents.[49]

The district began its pilot of the SBMMTES during the 2011/12 school year. The purpose of this pilot year was to learn from the successes and challenges that arose in order to position the system to productively scale-up over time. The pilot field tested the use of teachers' individual

---

Administrators who evaluate teachers according to the Stull form often receive no or little training on how to observe and assess teachers according to Stull expectations.

[46] The classroom observation measure incorporated teacher self-assessments and lesson planning activities, the actual classroom observations and pre- and post-observation conferences between teachers and observers.

[47] In contrast, the Danielson Framework for Teaching has 4 domains (1 Planning and Preparation, 2 Classroom Environment, 3 Instruction, and 4 Professional Responsibilities) divided into 22 components (that get tracked by observers), and 76 smaller elements. Each component defines a distinct aspect of a domain; two to five elements describe a specific feature of a component (Danielson Group, 2012). The full DFT standard-component-element schematic is displayed in the appendix.

[48] Traditionally, the *content validity* of an instrument has been assessed via content review by subject-matter experts (Kane, 2001); given the extensive history and usage of the Danielson Framework for Teaching (see section 3) and the extensive stakeholder review and adaptation process carried out in the study district throughout 2010/11, the content validity of the DFT is not examined here.

[49] The validity of these two externally-developed measures is not the focus of this paper.

growth planning activities, the classroom observation cycle based on the new DFT-aligned protocols, and the online platform for teachers and administrators to report observation notes and ratings. The pilot also field tested stakeholder surveys of students and parents[50] for a subset of the participating teachers. In addition, the district provided school-wide and individual value-added scores to all teachers in the district in 2011/12; only those teachers who taught in tested subjects received value-added scores.[51] (Sections 4 and 5 provide additional information about these measures and the analysis samples used in this paper.)

## 3 Discussion of Related Literature

Validity "refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA & NCME, 1999: 9). Here the "test scores" being evaluated are the DFT-based observation ratings for participating pilot teachers. Validity is not a property of the instrument itself, however — assessments themselves are neither valid nor invalid — instead validation involves an "integrated evaluative judgment" of the ways that results are interpreted and used in relation to the intended purposes and uses of the scores (Messick, 1989).

The measurement field's view of validity has evolved over time, incorporating a variety of interrelated terms and concepts. As Brennan (2013) explains, much of the focus from 1950 through the 1980s centered upon *construct validity* — a more theoretical paradigm under which observed scores are viewed as representations of the unobservable latent trait under study. Assessment experts emphasized through the years that construct validity "is relevant when the tester accepts no existing measure as a definitive criterion (APA, AERA & NCME, 1966: 13) and "undergirds all score-based inferences" (Messick, 1989: 35). Since the 1980s, however, due in large part due to the writings of measurement scholar Michael Kane, the thinking around validation has shifted from this theoretical (construct-based) focus toward a more grounded argument- and claim-based approach for addressing the practical problems of validation.[52] In turn, "it has not been easy to formulate a general methodology for validation," explained Kane (2001: 339), since the validity argument "will typically involve different kinds of evidence relevant to the different parts of the interpretive argument." Today, then, validation is considered a matter of degree, a continuous iterative process that serves both to build the case for the use of

[50] The district received at least one completed parent survey for only 18 different elementary teachers and 23 secondary teachers (7% response rate among parents). Given this low response rate, the parent survey measure is not analyzed here.

[51] Plans for the SBMMTES have changed since the inception of the pilot. There is some uncertainty around the way in which results will be aggregated and/or combined for summative evaluation purposes, and it is unclear how teachers' contributions to student achievement will be measured in the eventual SBMMTES-based evaluations. Because of this uncertainty, the *consequential validity* of the DFT measure — that is, evaluating whether the desirable consequences of the measurement procedure outweigh the negative consequences of such use (Kane, 2001) — cannot yet be examined in detail.

[52] Alternatively, Brennan (2013: 76-77) explains that the "trinitarian" model of validity — parsing out content, criterion, and construct validity — "has been roundly criticized for decades as being either too narrow (content and criterion) or too abstract (construct)… In a sense, the principle virtue of the argument-based approach to validation compared with past approaches is that it puts claims 'front and center' where they belong."

the instrument and support improvements in its design, interpretation and analysis (Herman et al., 2011).

In his writings, Kane lays out several different types of relevant inferences, including scoring inferences and inferences related to generalization and extrapolation. Several recent studies have, in turn, explicitly applied Kane's argument-focused approach in attempts to validate new teacher effectiveness measures. For example, Hill and colleagues (2011) relied upon Kane's approach in their examination of the extent to which value-added scores correspond to other indicators of teacher and teaching quality, while Bell and colleagues (2012) carried out an argument-based analysis of the Classroom Assessment Scoring System (CLASS) observation protocol — articulating first an interpretive argument and then describing and analyzing the varied types of evidence that could be used to build a validity argument for the instrument. This paper adopts an argument-based approach modeled in large part on these two recent papers.

*The face validity of the Danielson Framework for Teaching*

In recent years, implementations of modified versions of the Danielson Framework for Teaching (FFT) have been studied in several locations, perhaps most notably Cincinnati and Chicago. Cincinnati launched its FFT-based Teacher Evaluation System (TES) in the 2000/01 school year, with an observation instrument that includes four domains (essentially aligned with the FFT),[53] while in 2008/09 the Chicago Public Schools (CPS) launched its two-year Excellence in Teaching Pilot, an effort comprised of twice-yearly observations of teachers (again using a modified version of the FFT) as well as pre- and post-observation conferences between the principal and the teacher to discuss evaluation results and teaching practice (similar to the study district's pilot process). These two FFT implementations yielded interview and survey findings regarding teacher and principal perceptions of the Danielson Framework's validity. For example, in their interviews with teachers during initial TES implementation, Milanowski and Heneman (2001) found that teachers tended to understand and accept the FFT's domains, standards and performance levels, which were seen as consistent with teachers' own views about good teaching.[54] Similarly, in Chicago's pilot, participating principals and teachers both reported that their new FFT-based conferences were more evidence-based, reflective, and improvement-focused than those they had engaged in previously (Sartain et al., 2011).

Researchers have also linked FFT usage to changes in teacher practice. In Cincinnati, Heneman and Milanowski (2003) found that, although teachers involved in the first two years of the TES reported overall neutral reactions to the new system, they also reported that they believed the evaluation process led them to align their teaching to student standards, become more organized, improve lesson planning, and improve their classroom management skills. More recently, Taylor and Tyler (2012) found that mid-career math teachers' participation in TES was associated with subsequent test score gains among the teacher's students in ensuing years. However, Kimball

[53] Today each Cincinnati teacher is typically evaluated every five years due to cost constraints. During the TES evaluation year teachers typically experience four classroom observations: three times by an assigned peer evaluator (an experienced teacher external to the school) and once by a site administrator (Kane et al., 2011).

[54] This finding was echoed in three districts studied by Kimball (2002), where participants reported that the new FFT-based evaluation systems contributed to quality dialogues about instruction between evaluators and teachers.

(2002) found that changes in teacher practice in the three large districts in his study were not characterized as deep or meaningful — more focused on classroom management and the nature of classroom interactions. But it is perhaps Chicago's pilot that is most relevant here, in that it focused on training, two observations of teacher practices using an FFT-based rubric, and conferences between teachers and principals. And according to Sartain and colleagues (2009, 2011), Chicago's pilot was considered relatively successful by participants, and the authors judged the pilot to be an overall improvement from the previous evaluation system.

*Convergence of FFT ratings and student test score outcomes*

Given that finding a "single perfect criterion of teaching effectiveness is a practical impossibility" (Kulik, 2001: 10), many recent teacher effectiveness studies have built their arguments in large part around evidence of *convergent validity* (Campbell & Fiske, 1959; Kane, 2006), evaluating results in terms of the correlation(s) with other (admittedly partial and imperfect) measures of effectiveness. In particular, much of the recent empirical research in this vein — including the Bell et al. (2012) and Hill et al. (2011) validity argument papers — has explored the relationship between classroom observation ratings, such as those recorded using the DFT rubric during the pilot year in this district, and estimates of teachers' contributions to their students' test score outcomes (commonly referred to as "value-added" scores). The general assumption underlying such analyses is essentially that, irrespective of any measurement and estimation challenges, the quality of the teacher's observed performance in the classroom, as rated by a trained observer, *should* have a statistically significant relationship with the teacher's contribution to the test outcomes of the teacher's students that year.

Perhaps most notable among the studies analyzing relationships between value-added scores and FFT-based classroom observation scores is the Gates Foundation's prominent three-year Measures of Effective Teaching (MET) project, which engaged a broad array of psychometricians and statisticians (and dedicated over $40 million) towards investigating better ways to identify and develop effective teaching. The MET project ultimately enrolled over 3,000 teachers in Charlotte, Denver, Hillsborough County (FL), Memphis, New York City, Dallas and Pittsburgh (Bill & Melinda Gates Foundation, 2011). The MET project relied on expert outside reviewers to observe a common set of videotaped lessons.[55] In the project's second research report, Kane and Staiger (2012) found that FFT scores were positively correlated with student achievement gains from the prior year or from a separate class section. Results from separate class sections from the same school year indicated that classroom observation-based FFT results — i.e., limited only to scores for domain 2 (Classroom Environment) and domain 3 (Instruction) that address aspects of teaching directly observable in the classroom — were correlated with underlying value-added (persistent differences in measured student achievement gains associated with a teacher) at levels of 0.19 in math and 0.11 in English Language Arts (ELA), with similar

[55] For the culminating MET study (Mihaly et al., 2013), each participating teacher was asked to video record four classroom periods for each subject they taught as part of the MET project. Secondary teachers were asked to record two lessons from each of two sections of classes. Elementary teachers teaching self-contained classes were asked to provide video recordings of four class periods of English language arts instruction (including reading if applicable) and video recordings of four class periods of mathematics instruction. Independent raters managed and trained by the Educational Testing Service (ETS), and who did not know or work with the teachers, scored each video recording using three different observation protocols. All video recordings were scored using the FFT.

correlations evident (0.18 in math and 0.11 in ELA) between teachers' classroom-observation-based FFT results and their underlying value added from the prior year (Kane & Staiger, 2012).[56]

The MET project released its culminating findings in January 2013. As in the prior MET report, Mihaly and colleagues (2013: 39) found that the "stable" components[57] of teachers' classroom observation-based FFT results and their value-added on state tests had low to moderate positive correlations: 0.27 in math and 0.28 in reading at the elementary level and 0.41 in math and 0.17 in reading at the middle school level, suggesting that each of the two measures "captures some distinct unique dimension of effective teaching."

In total, though, the MET project has shown that assessing teachers via FFT and value-added scores will yield different rankings, and there is no clear statistical procedure that can decide which measure should count more in teacher evaluations (Rothstein & Mathis, 2013). The lessons from the MET work cannot entirely speak to what will be found in real-life situations, in contexts in which districts are actually implementing new systems for current or eventual stakes. To date, however, it has been difficult to study the relationships between the FFT in practice (pilot or otherwise) and student test-based outcomes because most multiple-measure teacher evaluation systems are relatively new. In their two-year study of Chicago's pilot effort, Sartain and colleagues (2011) also focused primarily on domain 2 (Classroom Environment) and domain 3 (Instruction) of Chicago's FFT-based framework, as these domains were the focus of the pilot. To assess the relationship between the Chicago pilot's classroom observation ratings and value-added results, the authors regressed the teacher's value-added score for that year against each of the teacher's 20 component-level ratings from domains 2 and 3 (20 separate models), then tested whether the ratings explained a significant portion of the variation in value-added measures (via omnibus F-tests). Sartain and colleagues found that the relationship between FFT-based observation ratings and test score growth was statistically significant for all components; teachers with the lowest observation ratings tended to have the lowest value-added measures and value-added measures tended to increase as ratings increased (with comparable results in math and reading).

In Cincinnati, relationships between teachers' TES scores and their students' test score outcomes have been examined in multiple studies (Holtzapple, 2003; Milanowski, 2004a, 2004b; Kane et al., 2011). In his early analysis of teachers' 2001/02 TES scores "within a value-added framework," Milanowski (2004a: 34) correlated the sum of teachers' four TES domain ratings in math and reading (grades three through eight) and the differences between predicted and actual student achievement in those subjects that year, finding moderate positive correlations "for most

[56] These results draw primarily from middle school teachers, who are more likely to work with more than one group of students during a year; however Kane and Staiger (2012: 47) note that a subset of the MET study's elementary teachers "also specialized by subject and taught more than one section of students."

[57] In this case the "stable component" of the observation or student test-based measure represents the teacher's average performance over a longer period of time and multiple classes, "similar to a universe or true score in a measurement model or generalizability theory" (Mihaly et al. 2013: 13). The authors restricted their study sample to include only "self-contained" elementary and middle school teachers with multiple classrooms: for elementary teachers, multiple classrooms were defined as two self-contained classrooms in two consecutive years (and these teachers were separately measured on math and ELA tests as if they were two separate classrooms). For middle school teachers multiple classrooms were defined by two separate sections of the same subject from the same school year (ibid: 8).

grades in each subject tested." Combining correlations across grades within subjects, the average correlations were 0.32 for reading and 0.43 for math ($p<.01$). These relationships align with more recent studies of Cincinnati teachers' TES scores, which indicate that TES performance has implications for student learning. Similar to other studies, in their analyses of the relationships between TES scores and student test score outcomes, Kane and colleagues (2011) focus only on results from domain 2 (Creating an Environment for Learning, with 3 standards) and domain 3 (Teaching for Learning, with 5 standards), as these are based on data from classroom observations rather than other evidence. Results from these Cincinnati studies generally align with previously discussed findings involving VAM scores, with TES evaluators awarding higher ratings to teachers whose students experienced higher gains in student achievement. According to Kane and colleagues (2011), improving a teacher's overall average TES score by one point was associated with one-seventh of a standard deviation increase in reading achievement among that teacher's students (and one-tenth in math), and the difference between being assigned a top-TES-quartile versus bottom-TES-quartile teacher was associated with a seven percentile gain in reading (and six in math).

Overall these studies have found that FFT- and student test-based measures of teacher performance tend to be positively (but moderately) correlated (see table C1 for an overview), with a common explanation for these low to moderate correlations being that each measure contains distinct information about the teacher's underlying effectiveness.[58]

Table C1. Summary of results from recent studies examining relationships between FFT results and student-test-based measures of teacher effectiveness

| Authors | Context: Locale | Observational Measure(s) | Sample | Key (FFT focused) Findings |
|---|---|---|---|---|
| Kane & Staiger, 2012 | Research: MET project Rd 2 study | FFT (plus CLASS, PLATO, MQI) | 1333 teachers | FFT correlated with underlying VA (persistent differences), both in current year (different classrooms): 0.19 math, 0.11 ELA<br>And in prior year: 0.18 math, 0.11 ELA |
| Mihaly et al., 2013 | Research: MET project Rd 3 study | FFT (plus CLASS, PLATO, MQI) | Estimated at 1000+ teachers | Correlations between stable components of FFT and value-added results (different classrooms):<br>Elem 0.27 math 0.28 ELA<br>Middle 0.41 math 0.17 ELA |
| Kane et al., 2011 | In Practice: Cincinnati TES | Modified version of FFT (domains 2 & 3 studied) | 207 teachers | 1 SD increase in teacher's average (normalized) FFT-based TES score across domains 2 & 3 associated with test score increases of 0.09 SDs in math and 0.08 SDs in reading on tests that year |
| Milanowski, 2004a | In Practice: Cincinnati TES | Modified version of FFT | 212 teachers | Averaging correlations across grades within subjects, the sum of teachers' 4 TES domain ratings was significantly correlated ($p<.01$) with the differences in predicted and actual student achievement in reading (0.32) and math (0.43) |

[58] The question of how high correlations should be to support claims of convergent validity is generally unresolved, though. Hill and colleagues (2011) note that "many take correlations of roughly .60 as strong evidence for convergent validity" (p. 798).

| | | | | |
|---|---|---|---|---|
| Taylor & Tyler, 2012 | In Practice: Cincinnati TES | Modified version of FFT (domains 2 & 3 studied) | 105 teachers | Average mid-career math teacher's students score 0.11 SD higher on math tests in years after teacher's TES evaluation than in years prior; non-significant difference in evaluation year (0.05 SD) and no significant impacts found in reading |
| Sartain et al., 2011 | In Practice: Chicago Excellence in Teaching Pilot | Modified version of FFT (domains 2 & 3 studied) | 501 teachers | Significant relationships (via F tests) found between ratings on 10 FFT components in domains 2 & 3 and VA scores; VA scores tend to increase as FFT ratings increase (evidence of linear relationship) |

Certain empirical issues have also arisen across these studies, in particular related to concerns about leniency among raters, potential nonlinearities between the two types of measures, and the potential for biased assignment of students to teachers. Despite extensive training and detailed rubrics provided to evaluators, evidence of leniency bias was evident in Cincinnati's TES system, where over 90 percent of teachers received a final rating that placed them in one of the top two categories (Taylor & Tyler, 2012), as well as in Chicago's pilot, where principals were more likely than external observers to rate teachers in the highest category (Sartain et al., 2011).[59] Given these results,[60] and since FFT-based protocols use an ordinal scale and are generally focused on cataloging practices (not student outcomes), the resulting scores are not necessarily linear functions of student test growth. In particular, Kane and colleagues (2011) found evidence of nonlinear relationships and asymmetry in the distributions of TES observation scores. Moreover, omitted variable bias is an almost universal empirical concern; in this context the estimated magnitudes of the relationships between the two types of measures may be sensitive to the nonrandom assignment of students to teachers — for example, the level of social cohesion among students may independently affect scores on both measures (Kane et al., 2011).

Such empirical concerns are noteworthy but generally beyond the scope of this paper. Here I do not seek to develop, refine or optimize any measure of teacher effectiveness. Instead I seek to examine the degree to which the DFT-based observation scores from a pilot implementation converge with the two other teacher effectiveness measures that were pilot tested in the district in 2011/12. To the latter end, this paper represents a more applied version of the MET studies.

[59] This does not imply that either observer was "right," however. Sartain and colleagues (2011) explained that, while principals used the Distinguished rating more often, these Distinguished teachers had higher value-added measures than the teachers the principals rated as Proficient, suggesting to the authors "either that principals are correctly identifying Distinguished practice or that they used historical knowledge of the teacher (unknown to the observer and outside of the evidence of the classroom observation) to form a better picture of teacher effectiveness" (p 16). When Sartain and colleagues accounted for teachers' previous evaluation ratings in their statistical models, much of the variation between principal and observer ratings disappeared.

[60] Other recent statewide pilots have yielded similar findings, with very high proportions of teachers deemed effective or better in Florida (97 percent), Georgia (94 percent), Michigan (98 percent) and Tennessee (98 percent) (Sawchuk, 2013). Although there may be a general expectation in designing observational rubrics that ratings will span rubric levels in ways approximating a normal distribution (McREL 2012), the distribution of "true" teacher performance is not necessarily definable.

*Student surveys as a measure of teacher effectiveness*

Soliciting students' perceptions of their teachers has a long history in the United States. In 1896, students in grades 2–8 in Sioux City, Iowa, provided input on effective teacher characteristics (Follman, 1995), and many U.S. colleges have used student ratings to evaluate teaching for decades. According to Kulik (2001), college students' ratings of their teachers tend to be highly correlated with both the students' exam scores and expert observations of the teacher. In their literature review on the topic, Renaud and Murray (2005: 930) concluded that "the weight of evidence from research is that (college) student ratings of teacher effectiveness validly reflect the skill or effectiveness of the instructor."

However, the literature on student feedback measures in K12 settings is less extensive. Some research findings offer promise regarding surveys' potential as a reliable and valid alternative measure of teacher effectiveness. Unlike school administrators and peer teachers, students have daily contact with their instructors, and have been found to be able to discriminate between effective and ineffective teachers, particularly at the secondary level (Ferguson, 2012; Follman, 1992; Worrell & Kuterbach, 2001). Students' ratings of teachers have also been found to be consistent, both from year-to-year (Aleamoni, 1999) and across different classrooms taught by the same teacher (Kane & Staiger, 2012).

In addition, Follman (1992, 1995) concluded that elementary and secondary student raters show no more evidence of leniency bias than adult raters, and other studies have found student ratings to be a moderate predictor of student achievement (Worrell & Kuterbach, 2001). In their study of nearly 2000 K12 students, Wilkerson and colleagues (2000) found that student ratings were significantly more accurate in predicting student achievement in both reading and mathematics than were teachers' self-ratings or principals' summative ratings. Moreover, MET project researchers recently found that one class's ratings of a teacher predicted achievement gains in a separate class taught by the same teacher (Kane & Cantrell, 2010). In the project's culminating technical report, the stable component of teachers' average student survey scores were found to have correlations with value-added that ranged from .33 to .44 in math and from .29 to .50 in ELA; alternatively, the correlations between teachers' student survey results and their average score across the eight FFT components in domains 2 and 3 — the Classroom Management and Instruction domains observable in a classroom visit — ranged from .41 to .50 (Mihaly et al., 2013: 24).[61]

At the same time, research has identified a number of known and potential disadvantages in using student surveys in any high-stakes teacher evaluation environment. For example, Goe and colleagues (2008) emphasized that students are unqualified to rate teachers on matters involving curriculum, classroom management, or content knowledge. Also, the earliest age at which students can offer a reasoned teacher rating is unresolved (Follman, 1995), and findings regarding the correlation between student grades and instructor ratings are inconsistent (Aleamoni, 1999), with some evidence suggesting that students with a higher grade expectation

[61] As noted, the stable component represents the teacher's average survey performance over multiple classes, with measurement and sampling error removed. Since I make no similar adjustment for within- or between-classroom noise in this paper, I do not expect to see correlations as high as those found in the culminating MET report by Mihaly and colleagues.

rate their teachers more favorably (Balch, 2012). It is also unclear how student ratings — generally only studied to date in low stakes pilot settings — would change in a high-stakes environment. "Mischievous adolescents given the opportunity to influence their teachers' compensation and careers via their survey responses may not answer honestly," Rothstein (2010: 7) pointed out. "Studies of zero-stakes student surveys can tell us little about how the students would respond if their teachers' careers were on the line." Such an environment might also impact teacher behavior as well, with some attempting to influence student ratings in unintended ways.

## 4 Data

Overall, the district's 2011/12 pilot was originally scaled to include approximately 700 volunteering teachers across 100 schools, rated by their site administrators (primary observers) as well as by a secondary observer (predominantly central office or local/regional office administrators). However, in part because the district required that every volunteering teacher be in a school with a volunteering principal, only 562 of the initial teacher volunteers were trained to participate. Another approximately 25 percent of these teachers dropped out after training, for a variety of reasons, including the outspoken opposition of the local teachers union, layoffs, transfers between schools, or evolving concerns over the workload involved. Ultimately, during the pilot year 371 teachers were rated by any observer in any cycle.[62] During the pilot, 125 site administrators (principals and assistant principals) served as primary DFT raters, and 210 individuals were trained as second raters, with 167 entering at least one rating for a participating pilot teacher.

This study relies on end-of-year surveys submitted by participating pilot teachers, collected by our research team, as well as three sets of data collected by the district during the pilot year. The first district dataset includes participating teachers' DFT-based ratings by primary and secondary observers across all 63 DFT elements. The second dataset includes teachers' value-added scores, as well as their value-added "level" in each subject/grade combination in which they had tested students. The final dataset includes teachers' average rating from their students' classroom experience surveys. Table C3 displays summary statistics for participating teachers on each of the pilot measures, while figure C1 shows the distributions of teacher performance across the different measures. I provide more details on these data below.

### *End-of-year survey responses from participating pilot teachers*

Our research team emailed survey links to participating pilot teachers at the tail end of the pilot year, with some additional follow up the following summer. Of the 371 teachers who received at

[62] An additional 36 teachers started the pilot process by filling out a self-assessment, but were never observed, and the district believes that at least some of these teachers along with another 19 may have been observed during the pilot year, but ratings were never entered for them by a primary or secondary observer. Pilot-year attrition — nearly half of initial teacher volunteers, and approximately a third of those trained, had no ratings entered online by an observer — must be kept in mind when interpreting the results presented here.

least one focus element rating from an observer in 2011/12, 192 completed surveys, for a response rate of 52 percent. These 192 responding participating pilot teachers worked in 105 different schools in 2011/12 and generally represent a group of experienced, mostly elementary teachers (see tables C2 and C4). According to our surveys, during the pilot year responding participating teachers had taught in their current schools for an average of nine years, in the district for 13 years, and had been teaching for 15 years overall. (California Department of Education data indicate that in 2011/12 the average district teacher had 13.8 years of teaching experience.) As shown in table C2, most of the responding participating pilot teachers reported teaching in elementary schools and reported teaching either multiple subjects or English language arts (ELA). Given the relatively low response rate and the characteristics of the survey sample, results cannot be considered representative of all participating pilot teachers, particularly secondary educators with, for example, math and/or science backgrounds.[63]

Table C2. Descriptive information about responding pilot teachers from end-of-year surveys*

| | |
|---|---|
| Responding teachers (full pilot sample) | 192 (371) |
| Schools represented | 105 |
| Average experience | Respondent mean |
| Years in current school | 8.8 |
| Years in the district | 13.3** |
| Years as teacher/administrator | 15.1** |
| Grade Level of Current School | % respondents |
| Elementary | 51.7** |
| Middle | 26.3 |
| High | 21.9 |
| Subject area (current/previous) | % responses |
| Elementary-Multiple Subjects | 27.6 |
| English Language Arts | 15.7 |
| Social Studies | 10.6 |
| Science | 10.6 |
| Math | 12.2 |
| Special Education | 4.0 |
| Arts | 5.6 |
| Physical Education | 5.0 |

* The end-of-pilot-year survey response rate for participating teachers was 52% (192 of 371).

** State data for 2011/12 indicate that district teachers' average years of overall service is 13.8 years and their average experience in the district is 12.9 years, and also indicate that 65.1 percent of the district's public schools are elementary schools. (Source: CDE Dataquest: http://dq.cde.ca.gov/dataquest/)

To examine teachers' perceptions of face validity, I focus on the proportion of respondents who agreed that the DFT standards on which they were rated "reflect my definition of effective teaching" and *also* agreed that formal DFT-based classroom observations "can provide an accurate assessment of my performance."

[63] There is also a degree of selection bias in the respondent sample. Results from our case site interviews and surveys indicated that volunteering teachers and administrators tended to believe strongly in the underlying tenets of instructional observation and feedback, and not only did survey respondents persevere through a pilot process that many considered burdensome, but they also completed an end-of-year survey about their experience. Many other teachers dropped out or didn't complete surveys, and their perspectives are absent here. Thus a certain amount of positive bias should be assumed within this survey sample, and any negative perceptions might understate actual difficulties.

*District Framework for Teaching (DFT) Ratings*

As noted, the District Framework for Teaching (DFT) includes five standards (Planning and Preparation, Classroom Environment, Delivery of Instruction, Additional Professional Responsibilities, and Professional Growth), 19 components (two to five per standard), and 63 elements (two to four per component). The full standard-component-element schematic for the pilot year is displayed in the appendix. The district's piloted DFT-based rubric for rating teacher performance included four rating levels (Ineffective, Developing, Effective or Highly Effective), with descriptors defining performance on each element at each level.

After initial trainings, pilot participants provided feedback to the district about the difficulties in collecting evidence about all 63 elements. Based on this feedback, the district narrowed the original 63 elements into 19 "focus" elements (FEs) on which observers were asked to rate participating pilot teachers during the pilot year. These 19 FEs clustered into the following seven focus components:

- 1d Planning & Preparation: Designing coherent instruction (contains 4 FEs)
- 1e Planning & Preparation: Designing student assessment (1 FEs)
- 2b Classroom Environment: Establishing a culture for learning (2 FEs)
- 3b Delivery of Instruction: Using questioning and discussion techniques (3 FEs)
- 3c Delivery of Instruction: Structures to engage students in learning (3 FEs)
- 3d Delivery of Instruction: Using assessment in instruction to advance student learning (4 FEs)
- 5a Professional Growth: Reflecting on practice (2 FEs)

Pilot teachers were to be rated in two separate observations (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 teacher's supervising site administrator and the second rater, generally based off site).[64] The observation data provided for this study include, for each of the DFT focus elements assessed during the pilot, the teacher's rating (coded one to four), the observation cycle (coded one or two), and the rater who provided the score (primary or secondary).

The district has not yet released guidance on how teachers' final "summative" DFT rating should be calculated, and there are several ways to create a single aggregate score. For example, as in Strunk and colleagues (2013a), one might calculate an overall average score across all focus elements (and standards); such a score offers a holistic view of the teacher's total work that year and represents the score teachers will receive (and on which their final evaluation score will be based). Alternatively, if the focus is on exploring the validity of a classroom performance measure (as is the case in this paper), then one might examine ratings for only those aspects of

[64] During summer and fall of 2011, multiple cohorts of potential observers participated in a 32-hour, weeklong training on the DFT. Observers watched and took notes on videos, then tagged evidence from their notes and applied ratings to their observations. Tagged evidence and ratings were reviewed by training leaders and rated on an evidence rubric and scoring accuracy measures. Based on these results, the observer was deemed either "certified" or "preliminarily certified," and some were provided with additional support before attempting certification again. However, not all observers in the pilot were certified or preliminarily certified (approximately 10% didn't attain certification), and the district has continued to work with them towards certification.

teaching that are directly observable in the classroom. Several recent empirical analyses of teachers' Danielson FFT-based ratings have adopted this type of narrower validation approach (Kane et al., 2011; Kane & Staiger, 2012; Mihaly et al., 2013; Sartain et al., 2011), and I also do so here. For this analysis I calculated participants' classroom observation-based DFT score as his or her average (mean) rating across the standard 2 (Classroom Environment) and standard 3 (Delivery of Instruction) focus elements rated in his or her cycle 2 observation by the primary rater. In turn, this analysis sample is limited to the 240 pilot teachers with at least one standard 2 FE rating *and* at least one standard 3 FE rating from a primary observer in cycle 2.[65]

I focus on only primary observer ratings for several reasons. First, the district has decided that, at least in the year following its first-year pilot, multiple raters will no longer be part of the district's SBMMTES, so these primary observer ratings most align with future SBMMTES iterations. Also, many second observers did not enter ratings in cycle 2, with interviews and surveys suggesting that many of those who did simply entered identical ratings as primary observers, at times after collaborating to reach consensus.[66]

To calculate each pilot teacher's DFT score for this paper, his or her available FE ratings were first averaged within standard 2 (which contains 2 FEs) and within standard 3 (10 FEs), and the resulting standard 2 and standard 3 means were then averaged. The resulting DFT scores for this sample of 240 pilot teachers ranged from 1.50 to 3.85, with a mean of 2.78 and a standard deviation of 0.45 (see table C3).

*Value-added Scores*

The research center hired by the district generated value-added measures of teacher effectiveness for the district for all teachers in grades and subjects covered by California Standards Tests (CSTs). The center generated one-year and three-year average value-added (VA) scores for the teachers whenever feasible. For the study district's teacher-level model, value-added is measured in math in third through eighth grades, ELA in third through ninth grades, and for various secondary level subjects like Geometry, Physics, and US History. Students' CST scale scores are

[65] These 240 teachers represent 65 percent of the 371 teachers with at least one FE rating from one observer during the pilot year. Averaging across both observers in both cycles or in each cycle would have dramatically reduced the paper's analysis samples. Only 50 teachers were rated on all 19 FEs by both observers in cycle 1, 33 teachers were thusly rated in cycle 2, and only 16 teachers were rated on all 19 FEs by both observers in both cycles. Moreover, only 72 teachers were rated *at all* by both observers in both cycles. Paper B, focused on the reliability of DFT ratings, examined these smaller samples in detail; there inter-rater reliability (Cohen's kappa) in cycle 1 was found to be .71 (50 teachers) and was .76 in cycle 2 (33 teachers).

[66] Although the district initially conceptualized the two ratings from observers as independent, for most of the pilot year observers were afforded discretion in how they input teacher scores. In some cases only primary observers input scores, other observer teams conferred before entering identical sets of scores, and in other instances each observer input his or her ratings independently. When asked on end-of-year surveys to select the option that best represented how they entered formal observation scores online during cycle 2, 64 percent of responding primary and second raters said they entered teachers' scores separately after formal observations, while 36 percent indicated that on average they collaborated and entered a single set of agreed-upon scores. In theory, an agreed-upon rating from two observers is likely to be a more accurate (i.e., less biased or noisy) rating of teacher performance. In this context, though, with limited information about the scoring process, it was not feasible to limit my analysis to only such consensus ratings.

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 are included. The center's analysis includes 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 (Center, 2011).

In general, the value-added model measures the "classroom effect" — the contribution a teacher makes to his or her students' average CST achievement, controlling for prior student achievement and a range of student- and classroom-level characteristics (Strunk et al., 2013a). 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/classroom's "value-add"). The value-added model for secondary level subjects yields estimates for teachers whose students took different pretests in prior years, and the secondary value-added model includes a term to control for average differences in posttest scores between students who took different pretests. All value-added results are standardized to center around three.[67]

Each teacher in the covered subjects receives an overall single- and three-year value-added 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 (Center, 2011). For my analyses for research question 3 I use the continuous one-year aggregate value-added score covering the pilot year (2011/12).[68]
Table C3 shows the descriptive statistics for these two value-added measures for all the teachers who participated in the pilot, while figure C1 displays the score distributions. As expected, the means are around three (as noted above, value-added scores were re-centered to three, so anything under three indicates a negative coefficient on the teacher fixed effect in the equation) and scores range from 0 to 6. Both value-added score means for the pilot sample are slightly above three, however, suggesting that the pilot teachers are slightly more effective on average, as measured by value-added, than other teachers in the district (Strunk et al., 2013a). And notably, 106 pilot elementary teachers had aggregate one-year value-added scores in *both* ELA (mean 3.07) and Math (mean 3.23) for 2011/12, and these two "within-teacher" value-added scores were highly correlated ($\rho$=.71, p<.01).

[67] For the single-year teacher value-added estimates cited in this paper, a three-stage regression to calculate teachers' fixed effects is run separately for each combination of grade, subject, and year over four years of data. An Empirical Bayes shrinkage approach is applied to ensure that teachers with fewer students are not unfairly overrepresented among the highest- and lowest-value-added teachers (Center, 2011).

[68] To construct these aggregate measures, the research center 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.

*Surveys of students*

Surveys of students' classroom experience were implemented for a subsample of participating pilot teachers in 2011/12. In the prior year, these surveys were developed and field-tested by local university researchers "for the purpose of measuring practices and conditions that align with the (District Framework for Teaching)" (Phillips et al., 2011: 71). First, the surveys were revised extensively through multiple rounds of stakeholder feedback, followed by limited field testing in the district's schools receiving federal School Improvement Grant funding in spring 2011. The researchers conducted focus groups with teachers and observed and interviewed students as they responded to the surveys. Field test data indicated that students and teachers tended to agree about many of the dimensions of the classroom environment they shared, and significant relationships were found ($p<.10$) between survey results and teachers' prior-year value-added scores, with results less correlated with value-added in English than in Math (ibid).

During the pilot year, paper versions of student classroom experience surveys were distributed in June 2012 to 75 pilot schools. This distribution included 2,757 elementary (grades 4-5) student classroom experience surveys and 83,132 secondary (grades 6-12) student classroom experience surveys. Among elementary surveys, 1,655 were returned (for a 60% response rate) and 32,802 secondary surveys were returned (for a 39.5% response rate). The surveys feature a Likert-style response scale, designed to estimate students' level of agreement on a 1 (low) to 5 (high) scale. The elementary classroom experience survey featured 42 survey items, seeking students' level of agreement with such statements as "My teacher is ready for every lesson", "My teacher treats students fairly", and "Most students in my class pay attention when the teacher is talking." The secondary classroom experience surveys were administered in English, History, Math and Science classes, with all surveys asking 50 general pedagogical items (e.g., "My teacher explains things in ways I can understand" and "My teacher treats me with respect") as well as three to eight subject-specific items (e.g., "Our teacher shows us how math is used in everyday life" and "My teacher connects history with things that are happening now").

Spring 2012 classroom experience survey results were submitted for 45 elementary teachers and 80 secondary teachers who participated in the pilot. The number of 2011/12 surveys submitted per teacher ranged from 4 to 34 at the elementary grades (with a mean of 25) and ranged from 5 to 180 surveys per teacher at the secondary grades (with a mean of 84). All survey results linked to participating pilot teachers were used; no required minimum number of item responses or respondents was applied. To serve as a convergent criterion for triangulation in this paper, an average student survey rating was calculated for each teacher by first averaging across all item scores (1-5) in each submitted survey (after reverse coding item responses where needed) and then averaging those means across all surveys submitted for that teacher.

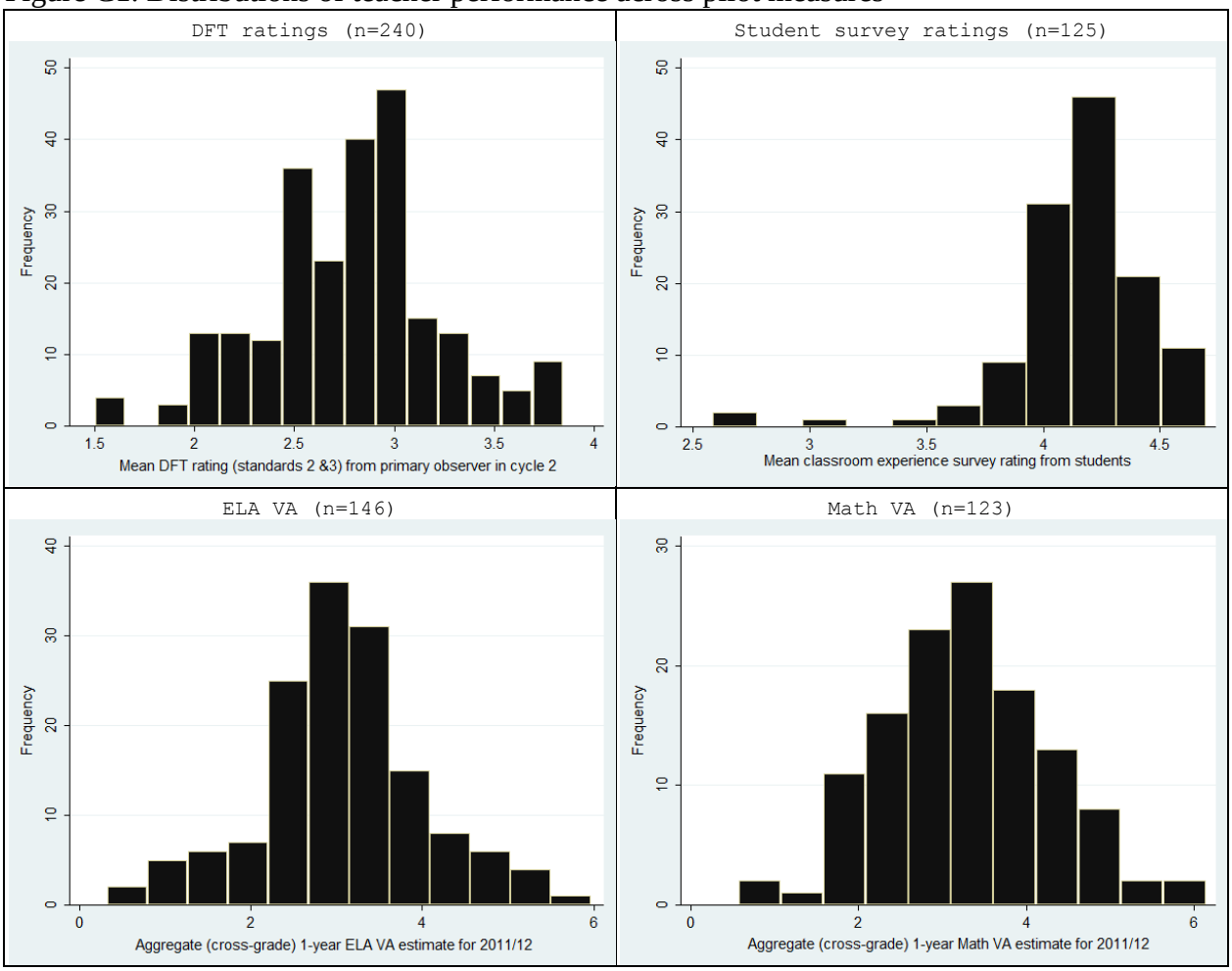
Table C3 displays summary statistics for pilot teachers' aggregate survey ratings in 2011/12. As shown, students tended to rate this sample of 125 pilot teachers quite highly (the mean rating was 4.16 with a standard deviation of 0.32 on a five-point scale), and figure 1 shows that the ratings distribution was negatively skewed (skewness= $^{-}$2.12), with most ratings clustered at the upper end of the scale.

Table C3. Summary statistics for 2011/12 teacher performance measures for pilot participants

| | Pilot teachers | mean | SD | min | max |
|---|---|---|---|---|---|
| DFT rating across standards 2 & 3 from primary observer in cycle 2 | 240 | 2.78 | 0.45 | 1.50 | 3.85 |
| Classroom experience survey rating from students | 125 | 4.16 | 0.32 | 2.59 | 4.70 |
| One-year ELA VA (across grades taught) | 146 | 3.09 | 0.96 | 0.33 | 5.97 |
| One-year Math VA (across grades taught) | 123 | 3.29 | 1.02 | 0.57 | 6.14 |

Note: 106 pilot elementary teachers had aggregate one-year VA scores in both ELA and Math ($\rho$=.71, p<.01).

Figure C1. Distributions of teacher performance across pilot measures



## 5 Analysis Samples & Methods

I rely on separate teacher samples to address each of my research questions. As shown in table C4, these samples featured higher proportions of non-White, female elementary teachers with ten or more years of experience than the district overall.[69] The mean California Academic Performance Index (API) scores across the samples' schools ranged from 733 to 797 in 2011/12, with no sample mean reaching the statewide API goal of 800. (In comparison, the district's API for the year was 744, while the overall statewide API was 788.)

In this paper I first explore the overall face validity of DFT classroom observation results as an indicator of teacher performance, based on teachers' perceptions from their participation in initial implementation in 2011/12. I tabulate responses using descriptive summary statistics and then used cross-tabulations to explore the ways in which teachers' perceptions relate to their average classroom observation-based DFT rating, their prior experience with similar activities at their school, their years of teaching experience, and the grade span taught (elementary/secondary). It's important to note that, given the strong selection bias inherent in the sample and the relatively low response rate among participating teachers (52%), the perceptions presented here cannot be generalized to the entire population of participating pilot teachers or the district as a whole. As shown in tables C2 and C4, results tend to reflect the perspectives of relatively experienced elementary teacher volunteers, and less so the perspectives of secondary educators with, for example, math and/or science backgrounds.

I next present validity evidence relevant to the different parts of the interpretive argument related to DFT scores. A first set of inferences relate to the scoring scale — if observers only ever used one or two points on the four-point scale, then one might question the degree to which the descriptions for the scale points were appropriate and whether the skewed scores reflected actual classroom quality and not scoring error. Results from paper B are relevant on this point. As noted there, the DFT scores awarded during the pilot year generally spanned all four of the scale's performance levels, although few Ineffective ratings were assigned and pilot teachers' scores tended to cluster at the higher end of the performance scale (Effective or Highly Effective) than at the lower end (Ineffective or Developing).[70] Not everyone got high scores, though, and the distribution generally reflected a range of teacher performance. Although over half of the scores given were 3's (Effective), the next most common rating was 2 (Developing), which tended to comprise 20 to 40 percent of scores across raters and cycles, indicating that there was room for growth and ceiling effects weren't overly problematic.

The argument for the scoring inference — in particular, generalizations made based on scores — can also be strengthened by analyses of bias that compare observers' scores for different groups

[69] Although not shown in table 4, virtually all participating pilot teachers (96.3%) had permanent (tenured) status.

[70] This evidence does not necessarily undermine the DFT scoring design, because it might be reasonable for actual teaching practice to be clustered around particular score points for this sample of teachers (as noted, discussions with site administrators suggested that these teachers tended to be experienced, hard working, and high performing), and the distribution of "true" teacher performance is not necessarily definable. However, if the scoring design is valid then one would expect to see additional converging evidence over time (e.g., from videos coded by expert/master observers) indicating that such skewed scores reflect actual instructional quality.

of teachers.[71] My analysis to address research question 2 thus explores whether significant differences were evident among different subgroups of teachers distinguished by characteristics that should (theoretically) be unrelated to their performance: the grade span they teach and their gender and ethnicity. Specifically, using the sample of 240 pilot teachers with at least one standard 2 FE rating *and* at least one standard 3 FE rating from a primary observer in cycle 2, I test whether mean DFT ratings differ significantly between groups of elementary and secondary teachers, male and female teachers, and White and non-White teachers.

Table C4. Descriptive statistics for analysis samples (2011/12 data)

| STUDY SAMPLES | Teachers | Grades taught | % White | % Female | % Elem. | % teaching 10+ yrs | % with masters/ doctorate | Mean school API |
|---|---|---|---|---|---|---|---|---|
| End-of –year participant surveys | 192 | 4-12 | 44.8 | 75.3 | 62.1 | 73.6 | 47.7 | 756 |
| DFT ratings | 240 | 4-12 | 40.2 | 78.6 | 70.5 | 79.0 | 41.5 | 779 |
| Student survey ratings | 125 | 4-12 | 38.2 | 70.7 | 50.4 | 70.7 | 49.6 | 733 |
| ELA value added (VA) | 146 | 4-12 | 42.5 | 70.6 | 84.5 | 77.4 | 43.8 | 773 |
| Math value added (VA) | 123 | 4-10 | 39.3 | 72.1 | 93.4 | 81.2 | 41.8 | 781 |
| DFT + Student Surveys | 79 | 4-12 | 37.2 | 74.4 | 52.6 | 79.5 | 46.2 | 751 |
| DFT + ELA VA | 92 | 4-9 | 45.7 | 73.9 | 85.9 | 80.4 | 42.4 | 784 |
| DFT + Math VA | 75 | 4-8 | 45.9 | 77.0 | 97.3 | 82.4 | 40.5 | 795 |
| Student Surveys + ELA VA | 69 | 4-10 | 39.1 | 69.6 | 81.2 | 75.4 | 44.9 | 762 |
| Student Surveys + Math VA | 53 | 4-8 | 36.5 | 73.1 | 88.5 | 75.0 | 40.4 | 782 |
| DFT + Student Surveys + ELA VA | 44 | 4-9 | 40.9 | 75.0 | 84.1 | 81.8 | 43.2 | 778 |
| DFT + Student Surveys + Math VA | 34 | 4-8 | 42.4 | 78.8 | 93.9 | 75.8 | 39.4 | 797 |
| ACROSS DISTRICT[1] | - | PK-12 | 41.1 | 68.4 | 65.1 | NA[2] | 41.4 | 744[3] |

[1] All district data were retrieved May 2013 from the California Department of Education (CDE) Dataquest system, online at http://dq.cde.ca.gov/dataquest/.

[2] CDE data for 2011/12 indicate that the average experience level among district teachers in 2011/12 was 13.8 years.

[3] This figure represents the district's 2011/12 API, not the average API of schools in the district.

Analyses for my final research question are designed to address "extrapolation" — connecting the teacher's score to some broader conception of his or her latent teaching ability (Bell et al., 2012; Kane, 2013). Exploring the validity of the extrapolation inference here invokes the basic social science tenet of triangulation — the extent to which teachers' classroom observation-based DFT scores converge with other contemporary measures of the same overall construct (teacher

[71] Kane (2001) also points out that an observed score should be "expected to generalize over various potential sources of irrelevant variance" (p. 333) such as raters and occasions, and that the existence of large rater effects "can suggest that generalization is too broad" (p. 331). Findings reported in paper B are relevant: Rater agreement was high in cycle 2 of the pilot (.76 across the 19 FEs) and my generalizability results indicated that the effects associated with observer group accounted for very little of the overall variation in pilot teachers' total DFT scores; observer ratings tended to be well calibrated at each cycle. Differences were evident, however, between teachers' observation ratings in cycle 1 versus cycle 2, with a quarter of a standard deviation increase in teachers' overall DFT rating (across 19 FEs), on average, between the two cycles. This is not necessarily surprising or problematic; increases between cycles 1 and 2 may indicate that teachers were able to align their practice with the DFT over the course of the year or that observers simply wanted to show increases in teacher effectiveness between cycles (and were more lenient in cycle 2). I cannot assess the degree to which either of these rationales holds true, but as noted I here focus my analysis on teachers end-of-year (cycle 2) scores. Ultimately, systematic differences among teachers accounted for 68 percent of the total variance in teachers' total (sum) scores across all 19 FEs. Thus, at least among the small sample of teachers rated twice by two observers, the majority of the variation in DFT scores was attributable to teachers rather than sources other than the teacher/classroom.

classroom performance during the pilot year).[72] My a priori expectation is that pilot teachers' mean DFT rating across standards 2 and 3 (representing the administrator's perception of the teacher's observed classroom performance against the DFT) will have statistically significant correlations with their scores from two contemporary (2011/12) teacher performance measures: their cross-grade value-added score (representing the teacher's estimated contribution to his or her students' standardized test outcomes) and their average student survey rating (representing his or her students' perceptions of their classroom experience that year).

First, I present correlations between mean classroom observation-based DFT scores and student survey results for the 79 pilot teachers (grades 4-12) with scores on both of those measures, then present the correlations between eligible pilot teachers' mean classroom observation-based DFT scores and their aggregate one-year ELA value-added scores (for the 92 eligible pilot teachers across grades 4-9) and their aggregate one-year math value-added scores (75 pilot teachers across grades 4-8). In order to comprehensively examine each of these relationships, I report separate Pearson product-moment correlations as well as Spearman rank order correlations (the latter makes no assumptions about the linearity of the relationship or the distribution of the data), and then I present partial correlations between classroom observation-based DFT ratings and value-added scores and student survey results and value-added scores (holding either survey or DFT ratings constant). As shown in table C4, 44 pilot teachers (grades 4-9) had aggregate DFT, student survey, and one-year ELA value-added scores for 2011/12, while 34 pilot teachers (grades 4-8) have aggregate DFT, student survey, and one-year math value-added scores.

Finally, I adopt regression approaches to separately assess the ability of participating pilot teachers' mean classroom observation-based DFT ratings alone, and then in combination with student survey results, to predict participating pilot teachers' value-added results for that year.[73] My first OLS analysis assesses whether adding an indicator improves the chance of identifying teachers who had larger student achievement gains that year, and I also create dummy (0/1) indicators for teachers in the top and bottom 10 percent of the aggregate one-year VA performance distributions in 2011/12 and examine via logistic regressions how well teachers' average classroom observation-based DFT and survey measures (separately and in combination) predict performance at these tails.[74] Such results have practical relevance because student surveys and DFT-based observations can be used repeatedly throughout a school year, while value-added results tend to be calculated after the school year is over. If the former measures can predict test outcomes on the latter, then their results might be used in a diagnostic or preemptive manner during the school year, for example having winter results focus additional supports to particular classrooms or teachers during the spring.

[72] As noted, the validity of inferences related to the uses, implications or consequences of DFT scores are not explored here, as these have not yet been clearly defined.

[73] I adopt this particular regression approach (regressing value added on classroom observation and survey scores) because it aligns with the recent research in this area, particularly the MET project) and not because it reflects any belief on my part about the legitimacy of any measure.

[74] This analysis also presents methodological options, and my approach differs from similar work carried out by Strunk and colleagues (2013a), who tested for significant differences between the overall DFT scores of pilot teachers ranked in the lowest of the study district's five value-added levels (far below average) and all other teachers, as well as between teachers who ranked in the highest of the five value-added levels (far above average) and all other teachers.

# 6 Results

*RQ1. To what extent did participating teachers report on end-of-year surveys that the DFT-based observations of their practice conducted during pilot implementation represented a valid measure of their effectiveness?*

Survey responses from participating pilot teachers suggested that their DFT ratings were indeed facially valid. Overall, at the end of the pilot year, 75 percent of responding teachers agreed that the DFT standards on which they were rated "reflect my definition of effective teaching" and *also* agreed that formal DFT-based classroom observations "can provide an accurate assessment of my performance." Again, though, these results may be heavily influenced by selection — as noted earlier, the sample of participants included in this paper are relatively more experienced elementary teachers who volunteered for the pilot, completed training, and were rated by an observer, *and then also* completed a survey about their experience — so the face validity claims made in this paper apply only to this select sample. Further research should be done to examine the face validity of the DFT among the full sample of pilot participants and the full population of district teachers.

There was also little variation in teachers' perceptions of face validity within this sample of survey respondents. First, the scores awarded didn't seem to matter, as the classroom observation ratings for those who agreed with the substance and accuracy of the DFT were not significantly different than those doubting its face validity on end-of-year surveys (p=.46). There were also no significant relationships between responding teachers' face validity beliefs and their years of experience (p=.32), gender (p=.50), or grade span taught (p=.49). Finally, whether pilot teachers' engaged in pre- and post-conferences and a formal evaluation during the prior school year also had no significant relationship (p=.44) with their perceptions of the face validity of the DFT. In general, this group of respondents felt that the observations of their practice conducted during the pilot year represented a valid measure of their effectiveness.

*RQ2. To what extent did participating pilot teachers' classroom observation-based DFT ratings correlate with theoretically-unrelated teacher-level variables, including the grade span taught by the teacher and his or her gender and ethnic background?*

As table C5 shows, although secondary teachers received slightly lower mean DFT ratings (across standards 2 and 3) from their primary observers in cycle 2 than did elementary teachers, the difference was not statistically significant (p=.16). Furthermore, no significant difference was evident between the mean DFT ratings for White and non-White teachers (p=.30).

There was, however, a statistically significant difference between the mean classroom observation-based DFT ratings of male and female pilot teachers (p<.01), with the 176 female teachers receiving a mean score that was 0.56 (pooled) standard deviations higher than the mean score for the 48 male pilot teachers. As shown in table C5, however, participating female pilot teachers also scored higher, on average, on the aggregate student survey and one-year value-added measures than did male pilot teachers, although the differences on the other pilot measures

were smaller and not statistically significant at traditional levels (except in the case of aggregate ELA value added). Moreover, the difference in mean DFT scores between male and female pilot teachers did not decrease and remained significant (at the .10 level or lower) in a series of OLS regressions where teachers' aggregate student survey and VA scores (either math or ELA) were included as covariates alongside gender. The discrepancy wasn't necessarily limited to pilot teachers, either. Across the entire district, among the more than 5000 teachers with value-added scores, female teachers had higher aggregate one-year VA scores in both ELA (by 0.14 SDs) and math ( by 0.09 SDs) in 2011/12 ($p<.01$).

Male and female pilot teachers differed in other ways as well. Although similar proportions of male and female pilot teachers had graduate degrees, 67 percent of female pilot teachers had ten or more years of teaching experience, compared to 52 percent of male pilot teachers. (Female pilot teachers had just over a year more teaching experience, on average, than male teachers, which amounts to a 0.36 SD gap in this sample.) This is distinction is worth noting — although the difference wasn't statistically significant at traditional levels ($p=.18$), pilot teachers with over ten years experience had average DFT ratings that were 0.15 standard deviations higher than those with under ten years of teaching experience in 2011/12.[75]

These qualifications — higher scores across the multiple pilot measures (and also among the district-wide value-added sample), coupled with more years teaching — suggest that the male-female DFT scoring discrepancy in 2011/12 may have been more related to actual classroom performance than bias.[76] In other words, although I find an association between gender and DFT rating, this does not necessarily reflect a lack of validity in the DFT. Rather, it may reflect the possibility that female teachers were more effective in the pilot year, as judged by multiple measures of effectiveness.

Table C5. Differences in mean classroom observation-based DFT scores (standards 2 and 3 only) from primary observers in cycle 2, by subgroups of pilot teachers

| Groups | Difference in Means (SD units) | p-value (one sided) | Pilot teachers |
|---|---|---|---|
| DFT: Elementary versus Secondary | 0.144 | 0.163 | 224 |
| DFT: White versus Non-White | -0.072 | 0.299 | 224 |
| DFT: Male versus Female | -0.564*** | 0.001 | 224 |
| On student surveys | -0.217 | 0.138 | 123 |
| On ELA value added | -0.369** | 0.021 | 146 |
| On Math value added | -0.131 | 0.259 | 122 |

* p<.10
** p<.05
*** p<.01

[75] This more experienced group of pilot teachers did not have higher average student survey or value-added scores.
[76] The panel data for this study do not include the demographic characteristics of the primary observers who assigned DFT scores, so empirical tests of particular gender biases are not feasible.

*RQ3. Among the participating teachers assessed on multiple teacher effectiveness measures during the pilot year, to what extent did their classroom observation-based DFT ratings correlate with contemporary measures, specifically their surveyed students' perceptions of their classroom experience and value-added measures of their effectiveness?*

Participating pilot teachers' classroom observation-based DFT ratings had moderate, statistically significant relationships with both their student survey ratings and their one-year aggregate ELA value added scores from 2011/12, with correlations ranging from .21 to .33 (see tables C6 and C7). These results are similar to those found in the culminating MET study (Mihaly et al., 2013), where, in a controlled research environment rather than a pilot implementation — with measurement and sampling error removed (which was not the case here) — the correlation between teachers' classroom observation-based FFT scores and their student survey ratings was .41 and the correlation with ELA value-added was .28. Unlike previous research, however, pilot teachers' DFT ratings had a lower correlation with their value-added scores in math.[77] This was not the case for student survey ratings, though, which had moderate correlations with math VA scores. I discuss several possible explanations for these results in the concluding section of this paper.

Table C6. Separate correlations between participating pilot teachers' aggregate DFT, student survey and one-year value-added ELA and math scores

| 2011/12 pilot performance measures | Pearson *r* | Spearman *ρ* | Teachers |
|---|---|---|---|
| DFT and ELA VA | .32*** | .33*** | 92 |
| DFT and Math VA | .16 | .23* | 75 |
| DFT and Student Surveys | .25** | .21* | 79 |
| Student Surveys and ELA VA | .29** | .25** | 69 |
| Student Surveys and Math VA | .39*** | .36*** | 53 |

* p<.10
** p<.05
*** p<.01

[77] My DFT-VA correlation results differ slightly from those presented in Strunk et al. (2013a). As noted, the DFT scores used in that paper were aggregated across a broader set of DFT focus elements and standards and represented a larger sample of pilot teachers (N=194). In their analysis for observation cycle 2, Strunk and colleagues (2013) found statistically-significant correlations between overall DFT ratings and both ELA VA (0.29, among 130 pilot teachers) and math VA (0.18, among 123 pilot teachers). These discrepancies — their parallel DFT-ELA VA correlation in cycle 2 is about .03 lower than the association presented here, and their DFT-Math VA correlation is about .02 higher — suggest that pilot teachers' non-classroom activities, specifically planning and reflection, had less of a relationship with their students' ELA test performance than did their classroom activities, while pilot teachers' non-classroom activities had a stronger relationship with students' math test performance than did their classroom activities. It is certainly possible that teacher planning and reflection influence students differently in different content areas (as the differing regression coefficients for DFT standards 1 and 5 in ELA and math in Strunk and colleagues' table 5 also suggest), but an in-depth exploration of such differences is beyond the scope of this paper.

Table C7. Partial correlations between pilot teachers' aggregate DFT/student survey results and their one-year value-added scores in ELA and math

| | ELA VA | Math VA |
|---|---|---|
| DFT Ratings | .31** | .05 |
| Student Surveys | .28* | .28 |

Note: These calculations only include the participating pilot teachers (44 in ELA and 34 in math) with scores on all three measures. The partial correlation between DFT ratings and VA scores is determined by holding student survey results constant; the partial correlation between student survey results and VA scores is determined by holding DFT ratings constant. The p-value for the partial correlation between student survey rating and math VA (*r*=.28) was .12.

* p<.10
** p<.05
*** p<.01

Parallel findings were evident from my initial OLS regression analyses. As table C8 shows, both classroom observation-based DFT ratings and student surveys ratings, separately and in combination, were statistically significant predictors of teachers whose students had larger ELA achievement gains that year.[78] Furthermore, among the sample of 44 pilot teachers with scores on all three measures, the proportion of explained variance (R-squared) in teachers' ELA value added increased — from 0.13 and 0.12 to 0.20 — when the two measures were included together in the OLS regression predicting ELA value added (table C8). Again, however, results were different in math, where only student survey ratings represented a significant predictor of math value added, although this was not the case (p=.11) among only the 34 teachers with all three types of scores.

Table C8. Regressing pilot teachers' one-year value-added (VA) results on their aggregate DFT and student survey ratings, separately and in combination

| | Predicting ELA 1-yr VA | | | Predicting Math 1-yr VA | | |
|---|---|---|---|---|---|---|
| OLS regression results | 1 DFT | 2 Survey | 3 Both | 1 DFT | 2 Survey | 3 Both |
| Mean DFT rating | 0.61*** (0.21) | - | 0.57** (0.27) | 0.37 (0.25) | - | 0.09 (0.31) |
| Mean student survey rating | - | 1.26** (0.56) | 1.02* (0.55) | - | 1.59*** (0.52) | 1.22 (0.77) |
| R-squared | 0.10 | 0.08 | 0.20 | 0.03 | 0.15 | 0.08 |
| Pilot teachers | 92 | 69 | 44 | 75 | 53 | 34 |
| Limited sample: Teachers with all 3 scores | 1 DFT | 2 Survey | 3 Both | 1 DFT | 2 Survey | 3 Both |
| Mean DFT rating | 0.70** (0.32) | - | 0.57** (0.27) | 0.16 (0.32) | - | 0.09 (0.31) |
| Mean student survey rating | - | 1.30** (0.60) | 1.02* (0.55) | - | 1.25 (0.76) | 1.22 (0.77) |
| R-squared | 0.13 | 0.12 | 0.20 | 0.01 | 0.08 | 0.08 |
| Pilot teachers | 44 | 44 | 44 | 34 | 34 | 34 |

Note: Robust standard errors [estimated via vce(robust)] are reported in parentheses.

* p<.10
** p<.05
*** p<.01

[78] Across pilot samples, DFT rating and student survey rating remain statistically significant predictors of ELA value added, at the .10 level at least, when gender, years experience, and/or an indicator for graduate degree (masters or doctorate) are added to the OLS model.

Table C9 indicates that pilot teachers' DFT and student survey ratings were both significant predictors of performance at the bottom end of the ELA value-added distribution. Within the pilot sample, those teachers in the bottom 10 percent of ELA VA in 2011/12 had mean DFT ratings that were 0.85 standard deviations lower than other pilot teachers (p=.01) and mean survey ratings that were 0.79 standard deviations lower (p=.02). And although the teachers in the top 10 percent of ELA VA in the pilot sample tended to have higher ratings on both measures, on average, than other pilot teachers, neither their mean DFT ratings (p=.24) nor their mean student survey ratings (p=.16) were significantly different from those of other pilot teachers. Within this sample, then, the two measures were better able to "predict" pilot teachers' ELA value-added performance at the low end of the distribution than at the high end.[79]

Similar to previous findings, pilot teachers' DFT ratings were not significant predictors of bottom- or top-10-percent math value-added performance among their pilot peers, although student survey ratings did identify performers at these tails (table C9). Specifically those teachers in the bottom 10 percent of math VA in the pilot sample had student survey ratings that were 1.3 standard deviations lower than other pilot teachers (p=.01), on average, while those in the top 10 percent of math VA in the pilot sample had survey ratings that were 0.8 standard deviations higher (p=.05).

Sample size is an important caveat, here, however. Although the top/bottom deciles of value added in this pilot sample each contained 12-15 teachers, not all of these pilot participants had DFT or survey scores, so my analyses of these tails relies in part on subgroups comprised of fewer than ten individuals.[80]

[79] Despite adopting differing methodologies, my results in this section (tables C9 and C10) parallel those found in Strunk et al. (2013a), who as noted tested for significant differences between the overall DFT scores of teachers ranked in the lowest of the district's five VA levels (far below average) and all other teachers, as well as between teachers who ranked in the highest VA level (far above average) and all other teachers. Their results (presented in their table A2) indicated that far below average ELA VA pilot teachers had an average overall DFT score that was significantly lower than all other pilot teachers, but no significant DFT difference was evident for far above average ELA VA pilot teachers, nor were significant DFT differences evident in math at the highest/lowest VA levels, on average. The authors noted that this was likely due in part to the very small proportions of teachers who fall into the extreme VA categories, and they found evidence of a more linear relationship between VA levels and overall DFT scores when they collapsed the five-level VA scale into three levels.

[80] Among the 14 teachers in the bottom ten percent of ELA value added in the pilot sample, 8 had mean DFT ratings and 8 had mean student survey ratings (4 had scores on both measures); among the 15 teachers in the top 10% of ELA value added in the pilot sample, 10 had mean DFT ratings and 8 had mean student survey ratings (5 had scores on both). Among the 12 teachers in the bottom 10% of Math VA in the pilot sample, 8 had mean DFT ratings and 4 had mean student survey ratings (2 had scores on both); among the 12 teachers in the top 10% of Math VA in the pilot sample, 10 had mean DFT ratings and 5 had mean student survey ratings (all 5 had scores on both).

Table C9. Differences (in SD units) between the mean DFT ratings for the top and bottom 10% of the value-added performance distribution versus the other 90% of pilot teachers

| | ELA VA | | Math VA | |
|---|---|---|---|---|
| Group | Bottom 10% | Top 10% | Bottom 10% | Top 10% |
| DFT rating | -0.85** | 0.23 | -0.23 | 0.17 |
| Student Surveys | -0.79** | 0.38 | -1.27*** | 0.79** |

Note: The figures displayed are in (pooled) SD units and were derived from t-tests of the differences in group scores, i.e., comparing the mean ratings for the top or bottom 10% against the mean ratings for all other pilot teachers. Although these 10% tails of the VA distributions each contained 12-15 pilot teachers, not all of these individuals had DFT or survey scores (see footnote 80 for details).

* p<.10
** p<.05
*** p<.01

Table C10 presents parallel results in terms of odds ratios from logistic regressions (rather than t-tests). Here the odds ratios for pilot teachers' DFT and student survey ratings (which are both continuous variables) essentially indicate how much the odds of ranking the top or bottom VA decile in the pilot sample increase or decrease when a DFT or survey rating increases by one unit (rating level in this context). So the first column of results in table C10 show that for a one-level *increase* in DFT rating, one would expect an approximate 84 percent *decrease* in the odds of ranking in the bottom ELA VA decile in the pilot sample (OR=0.16, p=.07). The odds decrease even further (by 98 percent) for a one-level increase in student survey rating (OR=0.02, p=.02). Independently, both measures are significant predictors of bottom-decile ELA value-added performance.

But here, as before, DFT rating is not a significant predictor of top-decile ELA value added or top- or bottom-decile math value added, while student surveys predict both high and low math value-added performance well, within the (admittedly limited) pilot sample. A one-level increase in student survey rating is associated with dramatically decreased odds of a bottom-10-percent Math VA ranking in the pilot sample (p<.01) and a very high increase in the odds of a top-10-percent Math VA ranking (p<.01). But in this context, incorporating a teacher's DFT rating alongside his or her student survey rating does not improve our ability to "predict" top or bottom value-added performance in ELA or Math.[81]

[81] I also carried out logit regressions to analyze how well pilot teachers' DFT and student survey ratings predicted teachers' rankings in the top 90%, top 50% and top 10% of the value-added distributions — specifically examining whether and how the coefficients and p-values shifted for each value-added rankings tier. Results generally paralleled those presented in tables C9 and C10, with DFT and survey ratings separately serving as significant predictors of top 90% and top 50% ELA value added (p<.10), but neither well predicting top 10% ELA value added. In math, student survey ratings by themselves served as a significant predictor of math value added (p<.10) across the three tiers (top 90%, top 50%, top 10%), while DFT ratings by themselves only served as a significant predictor of top-50% math value added (p<.05).

Table C10. Using pilot teachers' aggregate DFT and student survey ratings to predict the odds of top and bottom 10% of value-added performance

| | Bottom 10% ELA | | Bottom 10% Math | | Top 10% ELA | | Top 10% Math | |
|---|---|---|---|---|---|---|---|---|
| Logistic regressions | Odds Ratio | Model p-value | Odds Ratio | Model p-value | Odds Ratio | Model p-value | Odds Ratio | Model p-value |
| Mean DFT rating alone | 0.16* (0.16) | 0.07 | 0.61 (0.47) | 0.52 | 1.62 (0.70) | 0.27 | 1.44 (0.91) | 0.57 |
| Mean student survey rating alone | 0.02** (0.04) | 0.02 | 0.001*** (0.003) | 0.01 | 5.05 (9.51) | 0.39 | 56.23*** (84.08) | 0.01 |
| Both measures together | - | 0.02 | - | 0.02 | - | 0.71 | - | 0.04 |
| Mean DFT rating | 0.16 (0.18) | - | 2.56 (4.23) | - | 1.56 (0.96) | - | 0.63 (0.52) | - |
| Mean student survey rating | 0.04** (0.05) | - | 0.001* (0.006) | - | 2.10 (4.54) | - | 48.04** (74.82) | - |

Note: The p-values for each model result from Wald chi-squared tests of significance. Robust standard errors [estimated via vce(robust)] are reported in parentheses. Although these 10% tails of the VA distributions each contained 12-15 pilot teachers, not all of these individuals had DFT or survey scores (see footnote 30 for details). P-values do not change significantly when rankings are used in lieu of DFT and survey rankings.

* p<.10
** p<.05
*** p<.01

## 7 Limitations of this research

As noted earlier in this paper, this group of pilot participants was a self-selected, less heterogeneous group than the district's full teacher workforce. Teachers volunteered or were asked to participate in the pilot, with administrators reporting that participating teachers were a stronger, more committed and hard-working group, on average, than other teachers in their schools. In addition, due in part to transfers, layoffs, union opposition, and evolving perceptions of the workload involved, approximately a third of the teachers who participated in pilot trainings in fall 2011 never had a rating entered online by an observer. As a result, the findings presented here cannot be generalized beyond this particular pilot sample (generally experienced elementary teachers dedicated to the project).

Initial implementations such as this are also usually not the ideal contexts for validation studies. Participants and system developers tend to learn and make course corrections during the year, and start-up issues related to, for example, new trainings or technology may be confounded with technical issues related to the DFT measure. With greater training and increased practice, observers may in fact become more adept at observing, rating and recording teacher practice. The district could track this by, for example, examining over time whether there is greater variation in DFT ratings and if the relationships between teachers' value added and their DFT ratings become stronger or weaker.

# 8 Discussion

The group of participating pilot teachers who completed end-of-year surveys generally felt that the observations of their practice conducted during the pilot year represented a valid measure of their effectiveness. This aligns with other prior research on the face validity of the FFT (Kimball, 2002; Milanowski & Heneman, 2001; Sawyer, 2001) that found that during other early implementations, in places like Cincinnati (OH) and Reno (NV), participating teachers tended to report understanding and accepting the FFT's teaching domains, standards and outcomes.

Moreover, pilot teachers' classroom observation-based DFT ratings in 2011/12 were not related to their ethnicity or the grade span they taught, which makes sense, as these factors should theoretically be unrelated to individual performance. At the same time, however, there was a statistically significant difference between the average classroom observation-based DFT ratings received by male and female pilot teachers, with participating female teachers tending to score better against the DFT, on average. It's hard to determine the cause of this (theoretically-unrelated) gender discrepancy. Female pilot teachers actually scored higher across *all* of the 2011/12 pilot measures (not just the DFT) and had more teaching experience, on average, than male teachers. Although these findings are not necessarily indicative of gender bias, this is a measurement validity issue the district should continue to track in future years of DFT implementation.

Moderate correlations were evident between pilot teachers' classroom observation-based DFT ratings and their student survey ratings and ELA value-added scores for the 2011/12 year, a finding similar to other studies based in research settings (MET) and in practice (Cincinnati and Chicago). So from its applied setting, this paper offers some suggestive evidence that extrapolating from pilot teachers' classroom observation-based DFT ratings to their teaching performance is valid, at least in ELA. In turn, therefore, disaggregated DFT component or element ratings might well indicate particular areas for subsequent targeted support. (This connection was not a district focus during the pilot year.) At the same time, although the pilot teachers' in the bottom decile of ELA value added in the sample had significantly lower DFT ratings, the ratings for teachers in the top ELA value added decile (though slightly higher) were not significantly different from other pilot teachers.

As noted, there was little or no relationship between pilot teachers' classroom observation-based DFT ratings and their math VA scores in 2011/12, a finding that runs counter to the recent MET research. There may be a variety of explanations for this. First, over 93 percent of the pilot teachers with math value-added scores were elementary teachers, who tend to teach multiple subjects and are less likely to specialize in a content area than, for example, middle school teachers (who comprised a significant portion of the MET study sample). Notably, the pilot DFT data do not indicate the content area of the lesson observed, and clearly a rating derived from an ELA lesson would be more likely to correlate with students' ELA achievement than with their math achievement. Moving forward, the district would benefit from asking raters to note the content area of any elementary lesson observed, as this would enable a clearer examination of the relationships between administrator ratings and value added in that same content area (Strunk et al., 2013a). In addition, although we lack demographic data on observers, end-of-year survey results from pilot administrators indicated that relatively few previously worked as math

teachers. After elementary (multiple subjects), English language arts was the most common subject area previously taught by primary observers.[82] Administrators who have not worked as math teachers may only have a general knowledge of mathematics instruction and may, in turn, be less skilled in discriminating between levels of quality when observing it. Finally, it is possible that the DFT rubric, which is not subject area-specific, does not identify well teachers' impact on students' math achievement. Some recent research, particularly that underlying the Mathematical Quality of Instruction (MQI) observational instrument, 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 et al., 2012). This is another measurement validity issue the district should continue to track in future years of DFT implementation.

Another notable finding relates to the student surveys piloted. Although they were not the focus of this paper, teachers' aggregate ratings from their students' classroom experience surveys had moderate, statistically significant relationships with *all* other piloted measures (classroom observation-based DFT ratings and VA scores in both Math and ELA). This is promising, as gathering such information generally poses somewhat less of a logistical burden than observations and pre/post-conferences, and survey results tend to be more comprehensible to non-research audiences than value-added measures. The survey ratings clearly clustered at the higher end of the rating scale during this limited piloting, however, and questions remain about how results may shift when surveys are incorporated into any high-stakes professional decisions.

There is today some uncertainty in the district surrounding the use of the teacher performance measures piloted in 2011/12. Once further policy and accountability decisions have been made about the usage of DFT results, then the validity of inferences related to the implications or consequences of DFT ratings can be explored. Because the correlations between pilot teachers' aggregate DFT, student survey, and value-added scores were low to moderate, it is possible that the measures — arising separately from administrator perceptions, student perceptions, and statistical estimates of teacher contributions to student test outcomes — capture somewhat different dimensions of teacher performance (as suggested elsewhere). If these dimensions provide valuable information to teachers about their practice, then using multiple measures to assess teacher performance via such a triangulation approach makes sense (Strunk et al., 2013a). But from a practical perspective, given the limited information conveyed by value-added results and the logistical burden of observations/conferences, it may be more efficient to use value-added measures as an admittedly noisy pre-screening, no-stakes tool — perhaps by first identifying teachers who score significantly below or above the average value-added range as a first step toward planning which classrooms to observe and/or survey for diagnostic purposes the following year, a recent suggestion from both Baker and colleagues (2013) and Hill and colleagues (2011).

[82] The response rate for the end-of-year survey of participating primary raters was 53 percent (66 of 125), so these results aren't necessarily generalizable to the full pilot sample or district as a whole.

**References**


AERA, APA, NCME. (1966, 1999). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Aleamoni, L. ( 1999) Student ratings myths versus research facts from 1924 to 1998. Journal of Personnel Evaluation in Education 13(2), 153-166.

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

Bachman, L.F., Lynch, B.K. & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. Language Testing 12, 238–257.

Baker, B.D., Oluwole, J. & Green, P.C. III (2013) The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. Education Policy Analysis Archives 21(5), 1–68.

Balch, R. (2012). The validation of a student survey on teacher practice (Dissertation). Vanderbilt University.

Banchero, S. (2012, March 8). Teacher evaluations pose test for states. Wall Street Journal, p A2.

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.

Bill & Melinda Gates Foundation. (2010). Learning about teaching: Initial findings from the Measures of Effective Teaching project (Policy Brief).

Bill and Melinda Gates Foundation. (2011). Measures of Effective Teaching (MET). Retrieved 29 July 2011 from http://www.gatesfoundation.org/united-states/Pages/measures-of-effective-teaching-fact-sheet.aspx.

Bos, J., Sanchez, R., Tseng, F., Rayyes, N., Ortiz, L., & Sinicrope, C. (2012). Evaluation of Quality Teaching for English Learners (QTEL) Professional Development (NCEE 2012-4005). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

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

Brennan, R.L. (2013). Commentary on "Validating the Interpretations and Uses of Test Scores." Journal of Educational Measurement 50(1), 74–83.

Campbell, D.T. & Fiske, D.L. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 56(2), 81-105.

Clare, L. (2000). Using teachers' assignments as an indicator of classroom practice (CSE Technical Report, No. 532). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70(4): 213–220.

Danielson Group. (2012). Framework for teaching: Components of professional practice. Retrieved from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching.

Darling-Hammond, L. & Snyder, J. (2000). Authentic assessment of teaching in context. Teaching and Teacher Education 16, pp. 523-545.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. & Rothstein, J. (2012). Evaluating teacher evaluation. Phi Delta Kappan 93(6), pp. 8-15.

Datnow, A., Hubbard, L. & Mehan, H. (1998). Educational reform implementation: A co-constructed process. University of California Santa Cruz: Center for Research on Education, Diversity and Excellence (CREDE).

DeVellis, R.F. (1991). Scale Development. Newbury Park, NJ: Sage Publications.

Doyle, D. & Han, J.G. (2012). Measuring teacher effectiveness: A look "under the hood" of teacher evaluation in 10 sites. New York: 50CAN; New Haven, CT: ConnCAN; and Chapel Hill, NC: Public Impact.

Ferguson, R.F. (2012). Can student surveys measure teaching quality? Phi Delta Kappan 94(3), 24-28.

Fixsen, D.L., Naoom, S.F., Blase, K.A., Friedman, R.M. & Wallace, F. (2005). Implementation research: a synthesis of the literature. Tampa, FL: University of South Florida.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. High School Journal 75(1), 168–178.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. Child Study Journal 25(1), 57–78.

Goe, L., Bell, C. & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality.

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.

Herman, J. L., Heritage, M. & Goldschmidt, P. (2011). Developing and selecting assessments of student growth for use in teacher evaluation systems (extended version). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hill, H.C., Charalambous, C.Y. & Kraft, M.A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. Educational Researcher 41(2), 56–64.

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.

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.

Ho, A.D. & Kane, T.J. (2013). The reliability of classroom observations by school personnel. 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.

Honig, M.I. (2006). New directions in education policy implementation: Confronting complexity. Albany, NY: The State University of New York Press.

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

Kane, M.T. (2001). Current concerns in validity theory. Journal of Educational Measurement 38(4), 319–342.

Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), Educational Measurement (pp. 17–64). New York, NY: Praeger.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement 50(1), 1–73.

Kane, T.J. & Cantrell, S. (2010). Learning about teaching: Initial findings from the Measures of Effective Teaching project (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.

Kimball, S.M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. Journal for Personnel Evaluation in Education 16: 241-268.

Kirby, S. N., Berends, M. & Naftel, S. (2001). Implementation in new American schools: Four years into scale-up. Santa Monica, CA: RAND.

Klingner, J.K., Ahwee, S., Pilonieta, P. & Menendez, R. (2003). Barriers and facilitators in scaling up research-based practices. Exceptional Children 69(4): 411-429.

Knapp, M.S., Copland, M.A., & Talbert, J.E. (2003). Leading for learning; reflective tools for school and district leaders. Seattle, WA: Center for the Study of Teaching and Policy.

Kulik, J.A. (2001). Student ratings: Validity, utility and controversy. In M. Theall, P.C. Abrami and Mets, L.A. (Eds.) The Student Ratings Debate: Are They Valid? How Can We Best Use Them? (pp. 9-26). San Francisco: Jossey-Bass.

Loeb, S. & McEwan, P.J. (2006). An economic approach to education policy implementation. In M.I. Honig (Ed.), New directions in education policy implementation: Confronting complexity. Albany, NY: The State University of New York Press.

McCormick, K.M. & Brennan, S. (2001). Mentoring the new professional in interdisciplinary early childhood education: The Kentucky Teacher Internship Program. Topics in Early Childhood Special Education 21(3): 131-144.

McDonnell, L.M. & Elmore, R.F. (1987). Getting the job done: Alternative policy instruments. Educational Evaluation and Policy Analysis, 9(2): 133-152.

McLaughlin, M.W. (1987). Learning from experience: Lessons from policy implementation. Educational Evaluation and Policy Analysis 9(2): 171-178.

McREL. (2012). Teacher 2009-2010 report: Validation. North Carolina Department of Public Instruction. Retrieved July 2012.

Mead, S., Rotherham, A.J. & Brown, R. (2012). The hangover: Thinking about the unintended consequences of the nation's teacher evaluation binge. American Enterprise Institute: Special Report 2.

Medley, D.M. & Mitzel H.E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 247-328.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). New York: Macmillan.

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.T. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. Journal of Personnel Evaluation in Education 15(3): 193-212.

Milanowski, A.T. (2004a). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. Peabody Journal of Education 79(4): 33–53.

Milanowski, A.T. (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.

Milanowski, A.T. (2011). Validity research on teacher evaluation systems based on the Framework for Teaching. Paper presented at the April 2011 American Education Research Association annual meeting in New Orleans, LA.

Milanowski, A.T. & Heneman, H.G. III. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. Journal of Personnel Evaluation in Education 15(3), 193-212.

Newton, X. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: A generalizability analysis. *Studies in Educational Evaluation* 36, 1–13.

Ogawa, R.T. & Bossert, S.T. (1995). Leadership as an organizational quality. Educational Administration Quarterly 31(2): 224-243.

Pallant, J. (2007). SPSS survival manual: A step-by-step guide to data analysis using SPSS for Windows (3rd edition). New York: Open University Press.

Phillips, M. & Yamashita, K. (2011). The [Study] School District Pilot of Classroom and School Environment Surveys: A Technical Report Exploring Reliability and Validity in Nine School Improvement Grant Schools. University of California Los Angeles.

Renaud, R.D. & Murray, H.G. (2005). Factorial validity of student ratings of instruction. Research in Higher Education 46(8), 929-953.

Resnick, L. (1991). Shared cognition: Thinking as social practice. In L. Resnick, J. Levine & S. Teasley (Eds.), Perspectives on socially shared cognition (pp. 1-20). Washington, DC: American Psychological Association.

Rothstein, J. (2010). Review of Learning About Teaching. Boulder, CO: Great Lakes Center for Education Research and Practice

Rothstein, J. & Mathis, W.J. (2013). Review of "Have We Identified Effective Teachers?" and "A Composite Estimator of Effective Teaching": Culminating findings from the Measures of Effective Teaching Project. Boulder, CO: National Education Policy Center.

Sartain, L., Stoelinga, S. R. & Brown, E.R. (2009). Evaluation of the Excellence in Teaching Pilot: Year 1 Report to the Joyce Foundation. Chicago: Consortium on Chicago School Research.

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.

Sawaki, Y. (2010). Generalizability theory. In N.J. Salkind (Ed.). Encyclopedia of research design. Thousand Oaks, CA: SAGE Publications, Inc.

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

Sawyer, L. (2001). Revamping a teacher evaluation system. Educational Leadership 58(5), 44-47.

Sebastian, J. & Allensworth, E. (2012). The influence of principal leadership on classroom instruction and student learning: A study of mediated pathways to learning. Educational Administration Quarterly 48(4): 626-663.

Shavelson, R. & Webb, N. (1991). Generalizability theory: A primer. Newbury, CA: Sage Publications.

Shavelson, R.J., Webb, N.M. & Burstein, L. (1986). Measurement of teaching. In M.C. Wittrock (ed.), Handbook of Research on Teaching (3rd edition). New York: Macmillan.

Shavelson, R.J., Webb, N.M., & Rowley, G.M. (1989). Generalizability theory. American Psychologist 44(6), 922–932.

Smith, P.L. (1981). Gaining accuracy in generalizability theory: Using multiple designs. Journal of Educational Measurement 18(3), 147–154.

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

Spillane, J.P., Halverson, R. & Diamond, J.B. (2004). Towards a theory of leadership practice: a distributed perspective. Journal of Curriculum Studies 36(1): 3-34.

Strunk, K.O., Weinstein, T. & Makkonen, R. (2013a). New evidence on the relationship between multiple measures of teacher effectiveness in practice. Paper presented at the 2013 Association for Education Finance and Policy annual conference in New Orleans, LA.

Strunk, K.O., Weinstein, T. & Makkonen, R. (2013b). 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 in New Orleans, LA.

Supovitz, J., Sirinides, P. & May, H. (2010). How principals and peers influence teaching and learning. Educational Administration Quarterly 46: 31-56.

Taylor, E.S. & Tyler, J.H. (2012). The effect of evaluation on teacher performance. American Economic Review 102(7): 3628-3651.

The New Teacher Project (Teacher Project, 2009). Teacher hiring, transfer, and evaluation in [study district]. Final Report.

The Stull Act of 1971, California Education Code Section 44660-44665 (1971).

U.S. Department of Education. (USDOE, 2010). Guidance on Fiscal Year 2010 School Improvement Grants under Section 1003(g) of the Elementary and Secondary Education Act of 1965. Washington, DC. Retrieved online August 2011 from http://www2.ed.gov/programs/sif/sigguidance11012010.pdf.

U.S. Department of Education. (USDOE, 2011). Teacher Incentive Fund. Retrieved online July 2011 from http://www2.ed.gov/programs/teacherincentive/index.html.

Value-Added Research Center (Center, 2011). Academic Growth over Time: Technical Report on the [Study District] Teacher-Level Model, Academic Year 2010-2011. Wisconsin Center for Education Research, University of Wisconsin-Madison.

Veen, K., Zwart, R. & Meirink, J. (2011). What makes teacher professional development effective? A literature review. In M. Kooy & K. van Veen (Eds.), Teacher learning that matters: International perspectives (pp. 3–21). New York, NY: Routledge.

Webb N. M., & Shavelson, R.J. (2005). Generalizability theory: Overview. In B. S. Everitt and D. C. Howell (eds.), Encyclopedia of statistics in behavior science (vol. 2, 717-719). Chichester, England: John Wiley & Sons, Ltd.

Webb, N. M., Shavelson, R. J. & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. Handbook of Statistics 26, 81–124.

Wilkerson, D.J., Manatt, R. P., Rogers, M. A. & Maughanm R. (2000). Validation of student, principal, and self-ratings in 360° feedback (registered) for teacher evaluation. Journal of Personnel Evaluation in Education 14(2), 179-192.

Worrell, F.C. & Kuterbach, L.D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. Journal of Secondary Gifted Education 12(4), 236–247.

## Appendix: District Framework for Teaching (DFT) schematic for 2011/12 pilot year

| **Standard 1 Planning and Preparation** | **Standard 2: The Classroom Environment** | **Standard 3: Delivery of Instruction** | **Standard 4: Additional Professional Responsibilities** | **Standard 5: Professional Growth** |
|---|---|---|---|---|
| *1a: Demonstrating Knowledge of Content and Pedagogy* | *2a: Creating an Environment of Respect and Rapport* | *3a: Communicating with Students* | *4a: Maintaining Accurate Records* | **5a: Reflecting on Practice** |
| 1. Knowledge of Content and the Structure of the Discipline | 1. Teacher Interaction with Students | 1. Expectations for Learning | 1. Tracks Progress Towards Identified Learning Outcomes | 1. *Accurate Reflection* (focus element) |
| 2. Knowledge of Content-Related Pedagogy | 2. Student Interactions with One Another | 2. Directions and Procedures | 2. Tracks Completion of Student Assignments in Support of Student Learning | 2. *Use of Reflection to Inform Future Instruction* (focus element) |
| *1b: Demonstrating Knowledge of Students* | 3. Classroom Climate | 3. Explanations of Content | 3. Manages Non-instructional Records | 3. Selection of Professional Development Based on Reflection and Data |
| 1. Knowledge of Students, Skills, Knowledge, and Language Proficiency | **2b: Establishing a Culture for Learning** | 4. Use of Academic Language | 4. Submits Records on Time | 4. Implementation of New Learning from Professional Development |
| 2. Knowledge of How Children, Adolescents, or Adults Learn | 1. Importance of the Content | **3b: Using Questioning and Discussion Techniques** | *4b: Communicating with Families* | *5b: Participating in a Professional Community* |
| 3. Knowledge of Students' Special Needs | 2. *Expectations for Learning and Achievement* (focus element) | 1. *Quality and Purpose of Questions* (focus element) | 1. Information About the Instructional Program | 1. Collaboration with Colleagues |
| 4. Knowledge of Students' Interests and Cultural Heritage | 3. *Student Ownership of Their Work* (focus element) | 2. *Discussion Techniques* (focus element) | 2. Information About Individual Students | 2. Promotes a Culture of Professional Inquiry and Collaboration |
| *1c: Establishing Instructional Outcomes* | 4. Physical Environment | 3. *Student Participation* (focus element) | 3. Engagement of Families in the Instructional Program | |
| 1. Value, Sequence Alignment, and Clarity | *2c: Managing Classroom Procedures* | **3c: Structures to Engage Students in Learning** | *4c: Demonstrating Professionalism* | |
| 2. Suitability for Diverse Learners | 1. Management of Routines, Procedures, and Transitions | 1. *Standards-Based Projects, Activities and Assignments* (focus element) | 1. Ethical Conduct and Compliance with School, District, State, and Federal Regulations | |
| **1d: Designing Coherent Instruction** | 2. Management of Materials and Supplies | 2. *Purposeful and Productive Instructional Groups* (focus element) | 2. Advocacy/Intervention for Students | |
| 1. *Standards-Based Learning Activities* (focus element) | 3. Performance of Non-Instructional Duties | 3. Use of Available Instructional Materials, Technology and Resources | 3. Decision-Making | |

| | | | | |
|---|---|---|---|---|
| 2. *Instructional Materials, Technology, and Resources* (focus element) | 4. Management of Parent Leaders, other Volunteers and Paraprofessionals | 4. *Structure and Pacing* (focus element)* | | |
| 3. *Purposeful Instructional Groups* (focus element) | *2d: Managing Student Behavior* | **3d: Using Assessment in Instruction to Advance Student Learning** | | |
| 4. *Lesson and Unit Structure* (focus element) | 1. Expectations for Behavior | 1. *Assessment Criteria* (focus element) | | |
| **1e: Designing Student Assessment** | 2. Monitoring of Student Behavior | 2. *Monitoring of Student Learning* (focus element) | | |
| 1. Aligns with Instructional Outcomes | 3. Response to Student Behavior | 3. *Feedback to Students* (focus element) | | |
| 2. Criteria and Standards | | 4. *Student Self-Assessment and Monitoring of Progress* (focus element) | | |
| 3. Design of Formative Assessments | | *3e: Demonstrating Flexibility and Responsiveness* | | |
| 4. *Analysis and Use of Assessment Data for Planning* (focus element) | | 1. Responds and Adjusts to Meet Student Needs | | |
| | | 2. Persistence | | |

* Focus Component/Element