TeachingWorks
UNIVERSITY of MICHIGAN

working papers

Urban Teacher Center's Formative Assessments for Developing Teachers

Cara Jackson

TeachingWorks working papers are unpublished manuscripts that focus on the professional training of teachers. They involve analysis of data or literature and reflect "good thinking" – clear, systematic interrogation of issues critical in the field of teacher training.

These working papers are circulated to promote discussion. As such they are freely available to a broad audience interested in the study and improvement of ideas and practices in teacher education.

TeachingWorks working papers are subject to a blind review process that focuses on the relevance of the proposed work to pressing problems in teacher education, the transparency and relevance of the methods to the questions asked, as well as the quality of the writing. All submissions should be original.

The views expressed herein are those of the authors and do not necessarily reflect the views of the University of Michigan and/or TeachingWorks.

Urban Teacher Center's
Formative Assessments for Developing Teachers

Cara Jackson

Abstract: Given the importance of teacher quality and the limitations of using pre-hire characteristics to assess teaching potential, Urban Teacher Center has developed several formative assessments of its teacher candidates to ensure that they receive feedback to support continuous improvement as they develop their practice. UTC's formative assessment measures include coursework grades, ratings on a classroom observation rubric, and measures of professionalism and growth mindset. In this paper, we discuss the research related to each measure, evidence UTC has gathered on these assessments, and challenges that we face in implementing formative assessments. We end with key lessons learned in the first five years of program implementation.

INTRODUCTION

Extensive research has demonstrated that, of all in-school factors, the quality of the interactions between teachers and students has the greatest impact on student learning. Students placed with highly effective teachers for three years in a row significantly outscore those assigned less effective teachers (Sanders & Rivers, 1996). Teacher effectiveness has even been correlated with long-term student outcomes, including college enrollment and future earnings (Chetty, Friedman, & Rockoff, 2011).

Pre-screening of new teachers remains challenging. While easily observed measures of performance, such as academic credentials, are often highly valued in the hiring process, research finds the relationship between degrees or coursework and student achievement to be inconsistent and often statistically insignificant (Wayne & Youngs, 2003). Using data from Chicago and Texas, respectively, Aaronson et al. (2007) and Rivkin, Hanushek, and Kain (2005) both find that observable teacher characteristics explain very little of the variation in teacher effectiveness. Using data from North Carolina, Clotfelter, Ladd, and Vigdor (2007) find that test scores and licensure have positive effects on achievement, but find no statistically significant effect—and, in some cases, a negative effect—of having a master's degree. Thus, the research suggests that pre-hire characteristics are not strong predictors of teacher effectiveness.

While pre-hire data on performance might provide a stronger basis for predicting teacher effectiveness, currently, only about a quarter of traditional teacher preparation programs routinely gather information on the performance of their teacher candidates (Greenberg, McKee, & Walsh, 2013). This paper describes the efforts of an urban teacher residency program to design and implement formative assessments that validly capture knowledge, skills, and teaching practices likely to lead to effective teaching. As we've developed these formative measures, we have sought and continue to seek feedback on how to improve the formative assessment measures and address the challenges we've encountered during implementation. The purpose of this paper is to describe our formative assessment measures, the ways we've sought to evaluate the quality of these measures, and some challenges we've faced in implementing formative assessment measures.

We first provide an overview of Urban Teacher Center's program. We then describe each formative assessment measure in more detail, with attention to the evaluation of our measures and implementation challenges and with particular attention to our classroom observation rubric, which is the backbone of the entire program. We conclude by summarizing some key lessons, in an effort to provide insight to others seeking to embed formative assessment in teacher preparation programs.

UTC PROGRAM OVERVIEW

The Urban Teacher Center (UTC) was founded to provide high-need schools with effective teachers. We accomplish this through a multifaceted theory of action, one that combines 1) selective recruitment process, 2) intensive training and support through rigorous coursework, extensive clinical experience, and ongoing coaching and feedback, and 3) continual evaluation of performance.

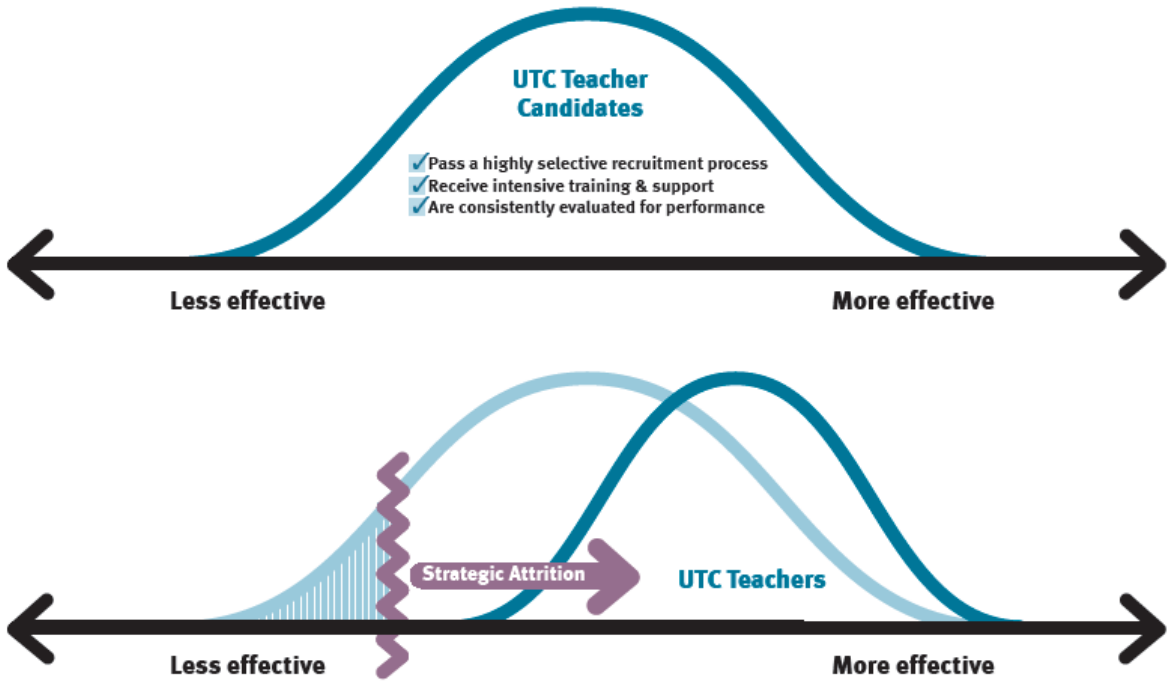

Figure 1: UTC Theory of Action

The bell curve on top encompasses the programmatic inputs: a strategic, selective recruitment process is follow by intensive training and support, which includes clinically-based graduate coursework, extensive classroom experience during the residency year, and a continuous cycle of improvement through regular coaching visits and feedback. Along the way, UTC systematically evaluates the performance of our participants, and we dismiss ineffective participants, typically after a probationary period. Thus, UTC's theory of action is grounded in the notion that a combination of selective recruitment, clinically-based training, ongoing support, and selective attrition will result in more effective teachers.

Since 2010, we have built an innovative, research-informed teacher preparation program from the ground up. Participants spend the first thirteen months, the residency year, working in classrooms with on-site coaching. At the same time, they take clinically-based graduate-level courses that introduce them to specific teaching practices and provide immediate opportunities to try those practices with students. Our clinical faculty, who serve as coaches, coursework instructors, or both, are key to the successful delivery of UTC's model.

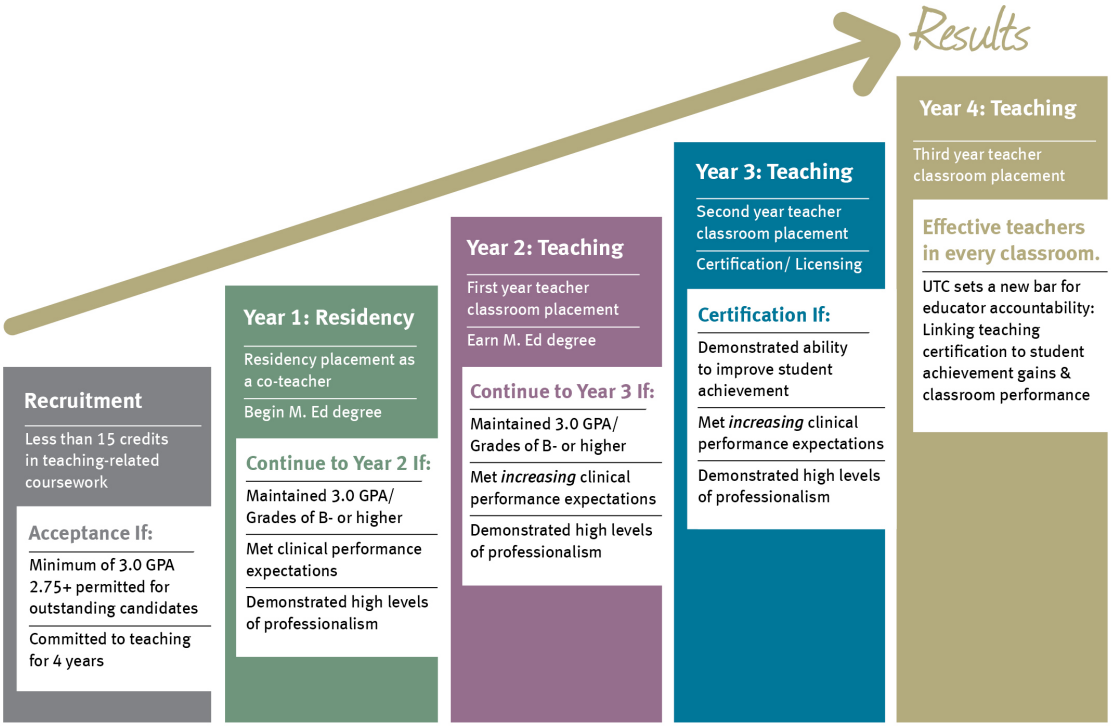

Figure 2: UTC's Program by Year

Participants spend the next three years working as full-time teachers of record in UTC partner schools, where they continue to receive regular, on-site coaching and support. At the end of the first year as teacher of record, successful participants earn their M.Ed. At the end of the following year (their third year with UTC), they become eligible for certification in both their content area and in special education. We provide preparation in both the content area and special education in order to enable our participants to meet the needs of diverse learners.

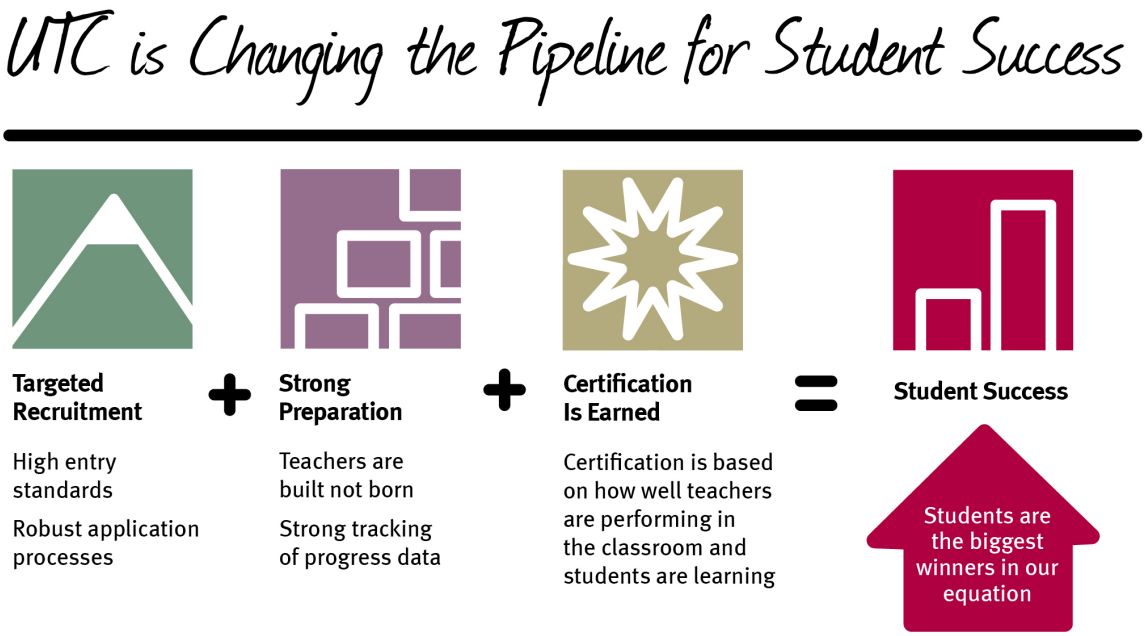

Figure 3: UTC Model Components Intended to Enhance Student Outcomes

This delayed decision around certification is key; it allows UTC to gather multiple years of on-the-job data to guide our final decisions about whom to recommend for certification. By the time our teacher candidates are eligible for certification, we have multiple indicators regarding our participants' ability to implement best practices and facilitate student learning in real classrooms, and we use these indicators to inform our decision regarding whether to recommend certification.

UTC's Formative Assessment Measures in Detail

In a radical departure from traditional teacher preparation programs, we gather data and provide feedback over a three-year period. These data guide the support and feedback that coaches provide in an effort to facilitate continual improvement in our participants. In addition, these data inform a hallmark of our work, UTC's Teacher Performance Assessment System, a multi-year and multi-measure system enacted prior to certification that helps us track the progress of every participant in our program. The Teacher Performance Assessment Systems is used determine who will advance from one stage to the next, thereby informing our deliberate approach to attrition.

UTC is one of a very small number of teacher preparation programs to take this step. In the absence of a broader accountability system for teacher education, we created our own system to assess whether our participants are equipped for the job prior to when they become fully certified. We view our approach as giving our participants an opportunity to succeed as a result of having multiple opportunities to practice teaching skills and receive feedback prior to evaluative use of the assessments. During each year of the program, teacher candidates must meet expectations across three domains to remain in good standing; the three domains are listed below.

1. **Courses grades:** Participants must pass each class with at least a B- (a requirement of our partner, Lesley University). Those who receive below a B- in the residency year are dismissed immediately, pending review. First-year teachers, who take a lighter course load in their first year of teaching, are held to the same grading standard, but they have an opportunity to retake a course, and dismissal decisions are not made until the end of the school year to minimize disruption to students. Participants must earn at least a B

average across the two years of coursework to earn a master's degree and advance to year three.

2. **Classroom practice:** Coaches conduct numerous classroom visits throughout the year and use UTC's Teacher Practice Rubric (TPR) to assess implementation of specific teaching practices. Focused (non-evaluative) observations are condensed coaching sessions which involve an observation and an immediate debrief. Comprehensive coaching cycles are formal and evaluative observations. The UTC performance and evaluation team reviews data from the Comprehensive Coaching Cycles at the end of each semester.
3. **Professionalism & Growth Mindset:** One strand of UTC's Teacher Practice Rubric is devoted to professional behaviors associated with effective teaching, such as receiving feedback openly and pursuing continued improvement in practice. Coaches and site teams gather data on these indicators for all participants and provide feedback on professionalism throughout the course of the year.

In addition to UTC's expectations, participants must meet state and local standards, such as passing licensure exams required by the state. UTC provides support to ensure that as many participants as possible meet expectations. Coaches and site directors will often intervene when they observe a resident or teacher falling behind, providing feedback and support to get them on track before a formal probation or dismissal conversation is needed.

Some attrition is a natural part of this process. We are very intentional about mid-program attrition, using performance indicators to quickly determine which teachers are not on track to becoming effective and to invest our energies in those who are. Nearly 80% of residents go on to become teachers of record. Of those who leave, many do so voluntarily, for personal reasons or because they have discovered that the challenges of teaching are not what they expected. About 40% of those who leave are formally exited through the assessment system outlined above.

UTC has drawn on existing research on measures of teaching quality to develop a system for ongoing assessment of the performance of our residents and teachers. We introduce the pieces of the resulting system in the sections that follow. Here we offer more detail about the research and thinking behind each of the measures, as well as the adjustments we have made—and continue to make—to ensure fair and consistent application of our assessment model.

Measure 1: Coursework Grades

In an effort to ensure a minimum quality of teaching, policymakers have often sought to set a floor for qualifications such as grade point average (GPA). GPA is considered an indicator of cognitive skills, which is positively associated with teacher effectiveness (Rockoff, Jacob, Kane, & Staiger, 2008). Although research suggests that overall undergraduate GPA is not significantly linked with teacher effectiveness (Kane, Rockoff, & Staiger, 2006), several studies have yielded promising if not consistently significant findings regarding the relationship between grades and grade point average (GPA) in teaching coursework and subsequent teacher performance.

For example, in a study of a university-based teacher preparation program, Henry et al. (2013) find that total GPA in upper division courses had a positive relationship with students' math achievement, though the relationship between GPA and reading achievement was not statistically significant. Guyton and Farokhi (1987) found GPA correlated significantly with the Teacher Certification Test score (a measure of content knowledge) as well as the Teacher Performance Assessment Instrument score. They also found that GPA in upper level courses, when students are taking education courses, had much stronger correlation with teaching performance than did GPA in the sophomore year; academic performance in education courses appeared to be a better predictor of teaching success than GPA in general. In a meta-analysis of 123 studies, D'Agostino and Powers (2009) find that performance in teacher preparation programs as measured by GPA was a better predictor of teaching skill than test scores. Wilson and Robinson (2012) reach similar conclusions based on a study of 1,800 teaching candidates.

In a more qualitative examination of coursework, the National Council on Teacher Quality scrutinized syllabi from teacher preparation program in an effort to assess coursework rigor. The authors argue that high grades and a lack of rigorous coursework in teacher preparation programs are largely a result of assignments that fail to develop critical skills and knowledge (Putnam, Greenberg, & Walsh, 2014). While the report has some methodological limitations, qualitative analysis of coursework assignments offers insight as to whether teacher preparation programs are giving participants an opportunity to practice teaching skills, and the presence of clear criteria for assignments provides an indicator of whether participants are receiving useful feedback on the skills practiced.

UTC Coursework

To ensure that coursework grades accurately reflect the extent to which participants are gaining the knowledge and skills required to teach effectively, our key assignments are clinically based and aligned to specific aspects of UTC's Teacher Practice Rubric. For example, in Emergent and Early Reading, one key assignment asks participants to assess students' knowledge of sight words, analyze assessment data, and reflect on the instructional implications of this data analysis, which is aligned with Strand B (Our Teachers are Diagnosticians) of the Teacher Practice Rubric. The key assignment for our Language Development in Children course requires participants to analyze discussion between students and explain how findings will impact future instructional decisions to extend students' linguistic development; this assignment reflects the teaching skills of Strand D (Fostering Academic Discourse).

These clinically based assignments are carefully structured to be aligned with our Teacher Practice Rubric, which delineates the skills teacher candidates need to learn. Assignments provide an opportunity for participants to receive feedback from their coursework instructor and their colleagues as they develop their skills (Ball & Forzani, 2010). Such assignments require collection of instructional data about student learning, identification of instructional challenges and how they were addressed, and reflection on practice, to foster meaningful conversations about the needs of our participants (Goe, Biggers, & Little, 2012). Subsequently, coursework grades reflect the quality of participants' work on assignments that are organized around core teaching practices, in keeping with recommendations from researchers (Grossman, Hammerness, & McDonald, 2009).

In addition, we have developed our syllabi and course materials to ensure alignment to the Common Core State Standards. We provide structured opportunities to examine and discuss the standards. For example, in Numbers and Operations and Algebraic Thinking II, required readings include resources from Progressions for the Common Core State Standards in Mathematics. The Progressions describe a specific topic or concept across a number of grade bands; they form the basis for the Common Core State Standards. Coursework provides ample opportunities to practice designing and delivering CCSS-aligned lessons. Residents deliver their lessons at their school site and bring evidence of student work back to coursework. In coursework, they reflect on their own performance and identify areas that need improvement. This exercise is repeated throughout the residency year.

Thus, grades on clinical and key assignments are formative assessments that are designed with the goal of improving our participants' teaching skills. Assignments are structured to sequentially provide participants with opportunities to develop content knowledge and pedagogical skills, apply knowledge and skills in teaching practice, and ultimately synthesize knowledge and skills in key assignments. Participants have opportunities to work on clinical assignments collaboratively and to receive feedback from peers as well as from course instructors. Throughout the course, participants receive feedback on their understanding and application of the underlying skills to foster continuous improvement. Consequently, course grades provide evidence of our participants' progress in developing teaching practices.

Course Instructors and the Implementation of Coursework Grades

Implementing coursework well requires deep and sustained support for our course instructors. To do so, UTC holds faculty institutes at the start of each semester and monthly meetings include discussion of strategies for course instruction, among other topics. Instructors

also have hour-long weekly calls by discipline area with a Lesley University faculty mentor to discuss course content, grading requirements, and evidence of participants' learning. We provide instructors with course curriculum and materials to support consistent implementation of our clinically-based graduate coursework. For example, UTC has developed sets of instructor notes for every course. The notes include pacing guides, suggested activities, and discussion questions.

To ensure that grading accurately captures participants' progress in developing teaching skills and to facilitate consistent grading across instructors, UTC syllabi include rubrics for key assignments that delineate specific criteria and provide descriptions of different levels of performance. These rubrics support rigor in our key assignments as they delineate clear performance standards. In addition, they provide a guide to course instructors to support them in providing constructive feedback to our participants.

In other efforts to support our internal commitment to programmatic accountability, we observe course instructors and provide formative feedback based on a rubric designed to capture key competencies. Instructors are observed twice during each course. Lead clinical faculty and members of the Curriculum and Professional Development team from UTC's national office assess coursework instructors the quality of planning and preparation as well as content and session development. Instructors are expected to serve as models for our participants, with well-planned, organized lessons. Additionally, UTC staff evaluate the timeliness and quality of communication and feedback provided by coursework instructors.

Validating Coursework Grades

Feedback from our participants indicates that coursework assignments are effectively promoting strong teaching practice. Based on our spring 2014 surveys of residents and teachers, about 90% of our participants agreed or strongly agreed that their math, literacy, and classroom management courses prepared them to teach effectively. In addition, we generally see a positive trend in the survey responses that suggest course quality is improving over time.

With coursework designed around best practices in the field and embedded in the classroom experience, our participants' course grades serve as a proxy for capturing their emerging competency in the classroom. Preliminary findings suggest that coursework grades predict participants' ratings on the Teacher Practice Rubric and student gains. For example, grades in Emergent and Early Literacy, which is taken during the residency year and has several assignments addressed at data literacy, had a significant, positive relationship with residents' year-long average on B1: Data Systems.[1] Although we only have student gains data on a small number of teachers, we find that grades in Emergent and Early Literacy have a significant, positive relationship with student gains in reading during the first year as teacher of record.[2]

Challenges with Coursework Grades as Formative Assessment

The first major challenge to using coursework grades as a formative assessment is ensuring consistency in grading practices and feedback on assignments. As noted above, UTC provides rubrics for key assignments in the course syllabi, and coursework instructors have weekly calls with Lesley University mentors as part of the ongoing training to support consistent grading and feedback. In addition, we have implemented systems to monitor the quality of coursework implementation. This includes the rubric for coursework instructors, which is used to assess the timeliness and quality of communication and feedback, as well as planning and preparation and content development. Finally, we have course evaluations to obtain stakeholder feedback; these evaluations include items to address whether instructors adequately explained grading methods and provided constructive feedback.

---

[1] This analysis was based on grades and TPR scores from 139 participants in cohorts 2012 and 2013. Specifically, we find a correlation of $r = .3$, $p < 0.01$.

[2] For this analysis, we only have data for cohort 2012 (n=22). Specifically, we find that a one point difference in course grade (e.g. an A compared to a B) is associated with a 0.4 standard deviation difference in student gains in reading.

The second major challenge associated with coursework grades is that many of the key assignments are clinically-based, so we need to identify schools and host teachers that are willing to allow implementation of these assignments and to accommodate assignment due dates. As noted by Davis and Boerst (2014), embedding teacher coursework in school settings requires building relationships and establishing creative structures with school principals and host teachers. Some of our school sites are less conducive to implementing coursework assignments than others, and our coursework instructors have had to work with participants to devise alternate routes of completing assignments.

The final major challenge associated with coursework grades is that the coursework is so tightly aligned with clinical experiences that our residents do not have the opportunity to repeat a course. Since they cannot receive a grade below a B-, this means that even average students may struggle to keep up.

Measure 2: Classroom Observation Ratings

Teacher practice rubrics are intended to offer fair, reliable judgments of the quality of teaching based on a shared understanding of what good teaching looks like. A variety of observational rubrics have been developed to capture different dimensions of quality of instruction, such as the extent to which teachers create and maintain an effective learning environment. Research suggests that trained evaluators, using rigorous teacher evaluation rubrics, can produce ratings of teaching practice that are significantly and positively correlated with student achievement growth (Kane, Taylor, Tyler, & Wooten, 2010; Milanowski, 2004). This correlation suggests that rubrics can unlock the black box of teaching and learning, accurately capturing the specific instructional practices that lead to enhanced student learning.

One of the greatest strengths of teacher practice rubrics is the potential to use the information diagnostically, giving teachers insight into how to improve their practice. Empirical evidence suggests that evaluations can support improvements in instructional practice. For example, Taylor and Tyler (2011) find evidence that quality observation-based evaluation and performance measures can improve mid-career teacher performance both during the period of evaluation and in subsequent years. In this regard, teacher practice rubrics are not merely a mechanism for evaluating effectiveness but can actually *promote* effectiveness. The observation of classroom practice by experienced educators, guided by a rubric that describes strong instructional practice, facilitates valuable conversations regarding novice teachers' work.

While observational rubrics *can* produce valuable data, they are not always implemented effectively. A widely reported study from TNTP in 2009 found that in districts using rubrics with a four-point scale, 94 percent of the teachers were awarded one of the top two ratings, and less than 1 percent were rated at the lowest level (Weisberg, Sexton, Mulhern, & Keeling, 2009). In a more recent update, TNTP found states and districts still rating most teachers in the highest categories, even after implementing new evaluation systems designed to produce a more realistic distribution of ratings (TNTP & Student Achievement Partners, 2013). A lack of variation in observational ratings is problematic in light of research that suggests teachers vary widely in their ability to generate student achievement gains (Aaronson, Barrow, & Sander, 2007; Rockoff, 2004; Rowan, Correnti, & Miller, 2002). Failure of observers to capture variation in teacher effectiveness undermines the potential for observation rubrics as both an evaluation tool and a diagnostic tool to improve teacher practice.

There are other potential limitations as well. Some research suggests that where observers do note differences among teachers, the variation in their ratings may, in some cases, reflect differences among the subjects taught or the type of lessons observed, rather than the overall quality of teaching, and inter-rater reliability may be low if observers do not have proper training and opportunities to calibrate their ratings with others. Even when observers are well trained, a single observation conducted by a single observer is likely to be somewhat unreliable. In the Measures of Effective Teaching project, researchers found that the reliability of observations increases from .51 to .58 when the same administrator observes a second lesson. By comparison, researchers report reliability increases from .51 to .67 if the second lesson is observed by a different administrator. Thus, the gain in reliability from adding another observer is

more than twice as large as the effect of adding another observation from the same observer (Ho & Kane, 2013).

UTC's Rubric Design and Coaching Cycles

Coaches conduct numerous classroom visits throughout the year and use UTC's Teacher Practice Rubric (TPR) to assess implementation of specific teaching practices. The TPR is consistent with the standards set by the Council of Chief State School Officers' Interstate Teacher Assessment and Support Consortium (CCSSO, 2011). It encompasses four sets of skills that new teachers must master in order to become effective. All of UTC's coursework and coaching support is built around these four areas of practice:

- Strand A: Build a productive and nurturing classroom environment
- Strand B: Serve as diagnosticians, using various forms of student data to guide instruction
- Strand C: Set precise goals and enact them
- Strand D: Foster academic conversations

Each strand includes multiple indicators; Strands A through D include a total of 19 indicators. Observers assign ratings for each indicator on a five-point scale, and the indicators are averaged to produce a single, summative score. Visits to assess performance on the TPR fall into two categories, described below.

Focused observations are condensed coaching sessions which involve an observation and an immediate debrief. The focused observation is 45 - 90 minutes in which the coach observes teacher practice for evidence of internalization, responsiveness to feedback, and solidification of teaching practices. The teacher observed is given immediate feedback that indicates next steps in their practice. During focused observations, UTC coaches may engage in real time coaching work to immediately address off-target teacher behaviors. Residents and second year teachers receive four focused observations during the school year; first year teachers receive six focused observations during the school year.

Comprehensive coaching cycles are formal and evaluative. In a comprehensive coaching cycle, the coach and participant have a conference both before the observation, to discuss what the coach will be looking for, and after the classroom observation, to discuss the coach's feedback. The post conference is an opportunity for coaches to acknowledge strengths and collaborate with participants to develop ideas for how to strengthen areas of practice. Comprehensive coaching cycles occur four times during the residency year, four more times during their first year of teaching, and twice during the second year of teaching.

In addition to focused and comprehensive coaching cycles, the development of participants' teaching practices is also supported by feedback on video submissions and classroom inter-visitations. These coaching components facilitate a shift to participants' investment in and ownership of growth in their teaching practices. For video submissions, the participant submits video footage of a lesson they are teaching to their coach, who reviews the video for evidence of whether the participant has followed through on previous coaching conversations, as well as solidification of the participant's instructional practices. The coach then provides constructive feedback for the participant to improve teaching practice. For inter-visitation, participants conduct site visits to other observe a classroom teacher who is exhibiting progress in identified areas. The participant has an opportunity to observe and converse with the teacher about the shifts in instructional practices and glean next steps for shifts in their own work.

Clinical Faculty's Implementation of Classroom Observation Ratings as Formative Assessment

Many of the limitations of teacher practice rubrics can be mitigated by investing in training observers and conducting appropriate oversight to ensure fidelity of their use. For example, raters participating in the Measures of Effective Teaching (MET) project underwent 17 to 25 hours of training and were required to rate a number of pre-scored videos and achieve a minimum level of agreement with the expert scores prior to certification. MET also monitored rater accuracy on

an ongoing basis, and those who failed calibration exercises could not score videos that day. Davis, Pool, & Mits-Cash (2000) note that inconsistent application of teacher assessment tools or abuse of instruments used to evaluate teachers can undermine the promise of teacher evaluation systems. Using multiple evaluators to observe a teacher reduces the risk that ratings are influenced by the personal (or subject-specific) biases of an individual evaluator.

UTC's approach to ensuring the quality of the data on teaching practice is multifaceted. To support consistency in communication to participants regarding what effective teaching looks like, we train and calibrate coaches before they begin classroom observations. UTC's Curriculum and Professional Development team holds faculty institutes at the start of each semester, during which coaches participate in norming and calibration exercises on the Teacher Practice Rubric to ensure consistency in use of the rubric and to assess inter-rater reliability. During monthly faculty meetings, coaches have in-depth discussions regarding classroom instruction and effective teaching practice. Throughout the semester, coaches occasionally conduct paired observations, which provides an opportunity to discuss evidence of teaching skills, ratings on the rubric, and what feedback they would provide to the participant.

Beyond these trainings, we've invested in an online system to improve consistency in how observers record their ratings and conduct analyses of coaches' data. Coaches use BloomBoard to schedule, manage and complete observations. BloomBoard allows coaches to share ratings and feedback, which facilitates greater transparency between the coach and the participant, as well as enabling UTC to monitor the data. We use BloomBoard reports to assess whether coaches are on track to complete their coaching cycles on time. In addition, we explore the data by coach to assess whether certain coaches have a tendency to rate more or less leniently. For example, in Figure 4, we see that coach number 4 rates lowest, on average, while coach number 13 has the highest average ratings.

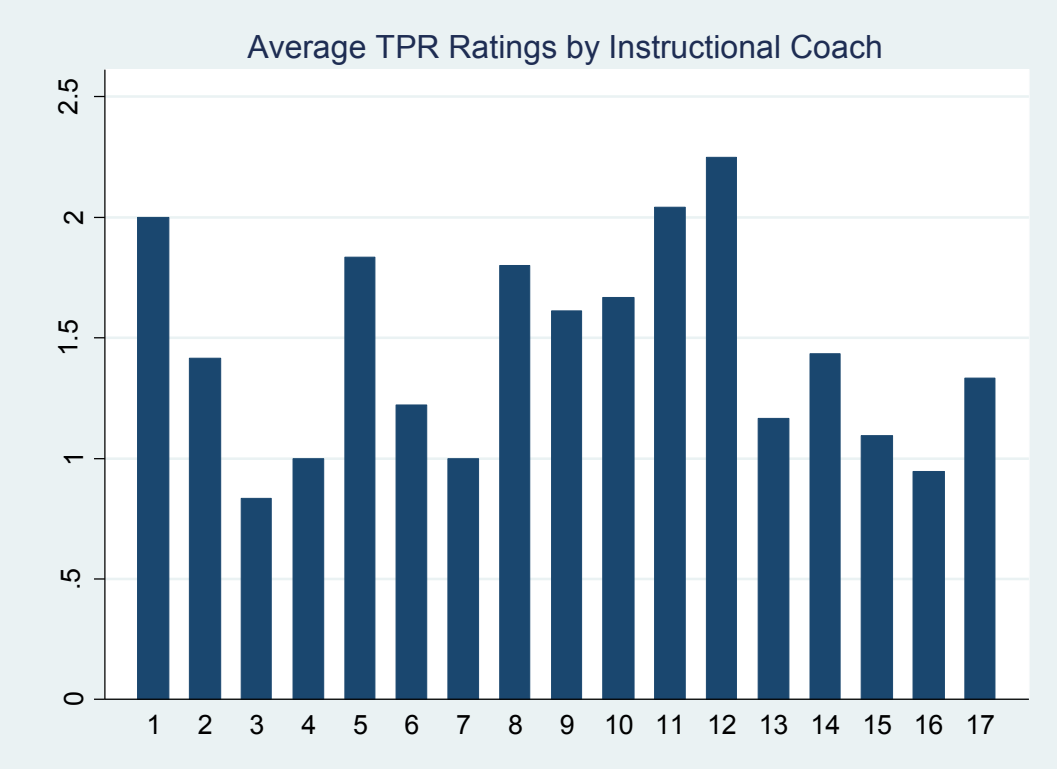

Figure 4: Average Teacher Practice Rubric Ratings for Residents, by Coach (SY 2014-15, October)

Since our coaches have relatively small caseloads, it is possible that a few outliers account for particularly high or low ratings. We use these data as a starting point for further analyses and conversations with others in the organization who are familiar with the idiosyncrasies of coaches' caseloads. Additionally, in keeping with the findings from the MET Project and other studies that suggest multiple observations increase data reliability, we draw on ratings from multiple observers and multiple observations for high-stakes decisions (Bill & Melinda Gates Foundation, 2012; Hill, Charalambous, & Kraft, 2012).

While ensuring consistency in coaches' ratings is obviously essential, it is also critical to attend to the quality of coaching feedback. To ensure that feedback is supporting our participants to become effective teachers, lead clinical faculty observe coaches and provide formative feedback on their performance based on a rubric that addresses key competencies. Coaches are assessed on content knowledge, timeliness and quality of communication and feedback, and quality of coaching.

Assessing Rubric Validity

UTC, in close partnership with Westat, has invested in assessing the validity of our Teacher Practice Rubric (TPR) to ensure its soundness as a measurement tool using data from SY 2012-2013. We have found that:

The TPR ratings are correlated with student achievement gains. Initial analyses by Westat indicate that rubric-generated ratings are positively correlated with student gains (r=0.51 for second year teachers and r=0.34 for first year teachers).[3] UTC confirmed these results with data from SY 2013-2014, finding that the average of a participant's TPR ratings across a school year is significantly related to student gains. Although these analyses are based on a relatively small sample of UTC participants for whom student gains data are available, the results suggest that the TPR accurately captured the variation in teaching practices that generate student achievement gains.

The TPR is internally consistent. Each of the rubric's strands contains three to six indicators that are averaged to produce a score for that strand. We would expect ratings of individual indicators to be largely consistent with each other, since they are intended to measure the same underlying ability or skill. Our evaluators found that ratings for all strands demonstrate strong levels of internal consistency (above .80, Cronbach's alpha).

The TPR ratings capture a range of practice. An effective rubric should differentiate among teachers at different levels of practice. We would expect to see substantial variation in ratings if a rubric is implemented well, particularly among novice teachers who enter teaching with different experiences and whose skills may develop at different rates. UTC's observational ratings do show substantial variation and do not cluster strongly in the middle or top categories (Figure 5). There is no indication of the leniency that TNTP reported finding in other teacher performance evaluation systems (Weisberg, Sexton, Mulhern, & Keeling, 2009).

[3] Our external partners, Education Analytics, obtained student gains estimated on the NWEA Measures of Academic Progress assessment for 20 year second-year teachers (cohort 2010) and 32 first-year teachers (cohort 2011). Education Analytics used a standard form of a value-added model estimated within a large data sample with one exception: due the small number of UTC participants, the parameters were not estimated, but instead calibrated from other data and then assumed to be accurate for students and teachers associated with UTC.

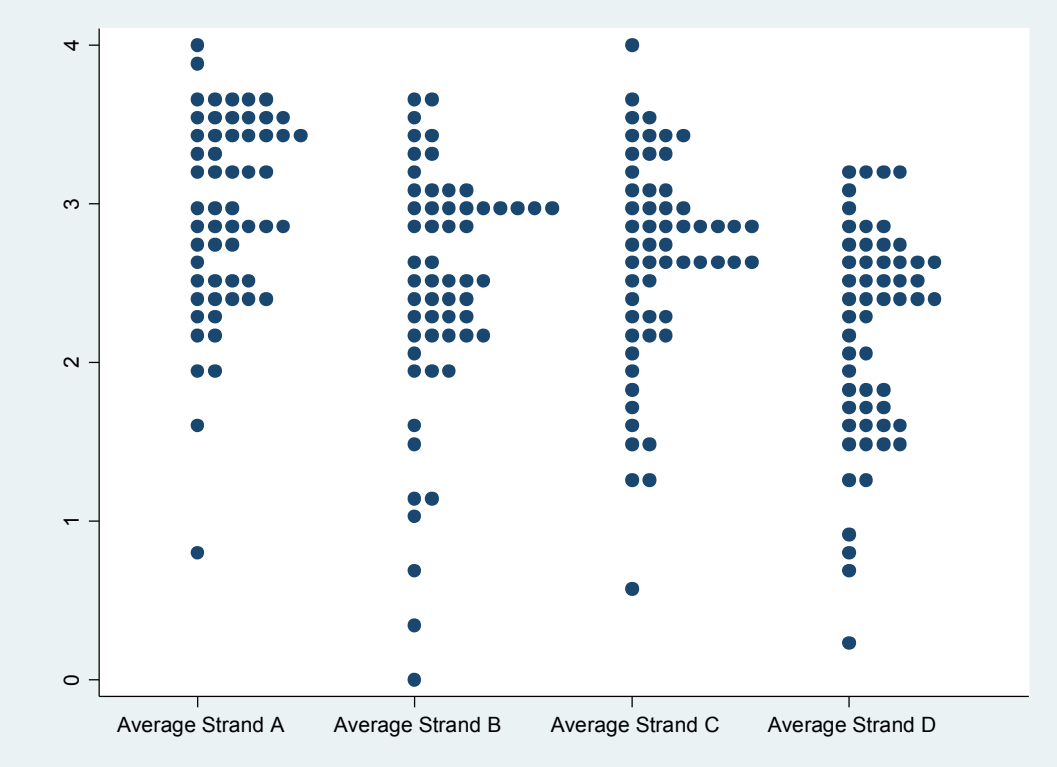

Figure 5: Distribution of Strand Ratings for First Year Teachers (Semester 2 of SY 2012-2013)

Ratings improve as teachers develop their practice. Westat's longitudinal analysis of TPR ratings finds that ratings of UTC novice teachers increase over the first few years of teaching, which is consistent with a considerable body of research indicating that, on average, teachers improve considerably in their first few years on the job. On average, estimates of teachers' value-added to student achievement gains increase in the first few years of teaching. Several studies have documented a leveling off of returns to experience. For example, Clotfelter, Ladd, and Vigdor (2006, 2007) find that about half the returns to experience occur within the first one or two years of experience.

Figure 6 displays the ratings of three UTC cohorts during the course of school year 2012-2013. Two patterns are evident. First, early career teachers improve their practice over the course of a school year; within each cohort, average ratings for the first observation are lower than average ratings for the last observation. Second, experienced teachers perform better on average than novice teachers; cohort 2010 (second year teachers) have the highest ratings on average, while cohort 2012 (candidates still in their residency year) have the lowest ratings on average.

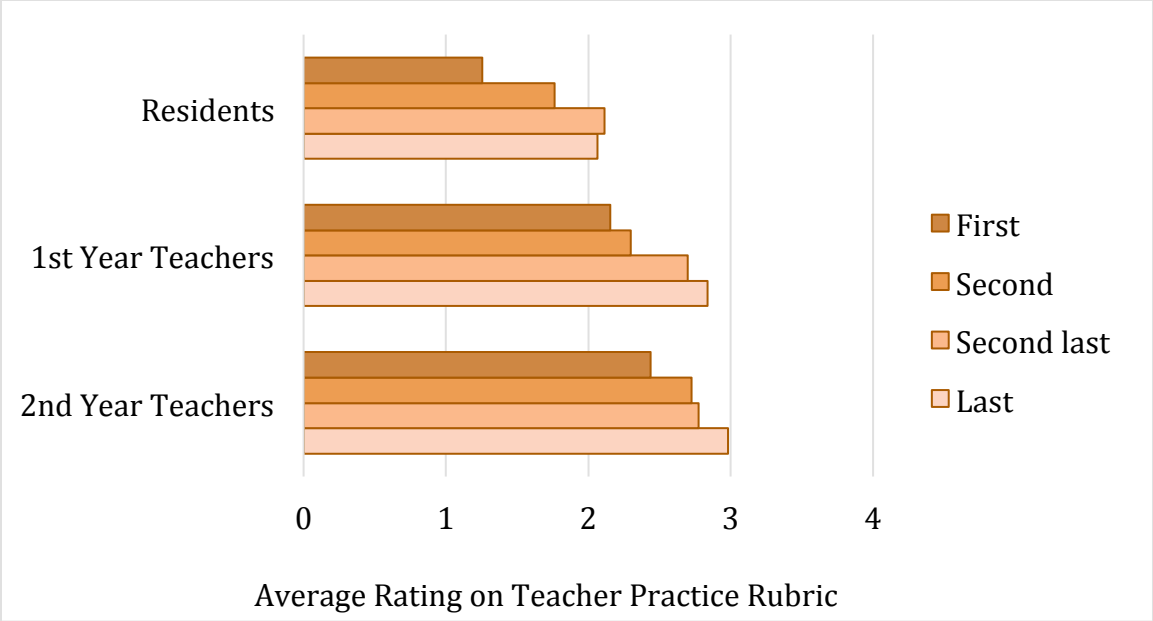

Figure 6: Teacher Practice Ratings by Year in Program, 2012-2013

Challenges with Classroom Observation Ratings as Formative Assessment

The first challenge with using classroom observation ratings as a formative assessment has to do with ensuring the quality of the data. The norming and calibration exercises that allow us to assess inter-rater reliability consumes a considerable amount of our time during clinical faculty institutes when conducted for all 19 indicators in Strands A through D of the Teacher Practice Rubric. In addition, ongoing monitoring of TPR ratings by coach is required to detect rater "drift." That is, we need to ensure that our coaches do not develop tendencies to rate more or less leniently. We also look for evidence that the coaches are applying the rubric as intended; if coaches are giving the same ratings to all participants, this would suggest a lack of fidelity in implementation. Responding to these rater issues can be resource-intensive. We currently devote one clinical faculty member's time to conducting paired observations with coaches that need additional training. Doing so requires a trade-off in that supporting coaches diverts resources away from supporting participants.

The second major challenge is related to the complexity of the coaching cycle. Strands A through D of the Teacher Practice Rubric includes 19 indicators, but residents and first-year teachers of record are not rated on all indicators. The complexity of the coaching cycle reflects an effort to tailor coaching to participants' year in the program. Participants are evaluated on indicators that capture the teaching skills they have learned in coursework and have had an opportunity to practice through key assignments. For those coaches who have a mix of residents, first-year teachers, and second-year teachers on their coaching roster, this complexity requires coaches to keep track of which indicators they need to rate given where the participants are in the program.

The third major challenge is that some school settings are not conducive to observing certain aspects of the rubric. Specifically, in Strand A one indicator focuses on whether the teacher "Arranges the physical environment to facilitate student engagement in learning and participation through visual, auditory, and movement opportunities." Some schools may place constraints on the teachers' creativity in arranging the physical environment. In Strand D, indicators focus on whether teachers are implementing strategies to foster academic conversations and encouraging conversations between students; some schools are philosophically opposed to student-centered discourse. Thus, we need to attend to the schools' approach to teacher autonomy and student voice when recruiting school partners.

Measure 3: Professionalism & Growth Mindset

Much of the research and national standards that mentions professionalism refers to a fairly broad vision of professionalism. For example, CAEP's standards for the accreditation of educator preparation programs include that the provider ensures that candidates demonstrate an understanding of the InTASC standards, including professional responsibility (CAEP, 2013). The professional responsibility standards in the Council of Chief State School Officers' Interstate Teacher Assessment and Support Consortium are listed below (CCSSO, 2011).

- Standard #9: Professional Learning and Ethical Practice. The teacher engages in ongoing professional learning and uses evidence to continually evaluate his/her practice, particularly the effects of his/her choices and actions on others (learners, families, other professionals, and the community), and adapts practice to meet the needs of each learner.
- Standard #10: Leadership and Collaboration. The teacher seeks appropriate leadership roles and opportunities to take responsibility for student learning, to collaborate with learners, families, colleagues, other school professionals, and community members to ensure learner growth, and to advance the profession.

Many teacher preparation programs have devised assessments of the dispositions of teacher candidates; however, dispositions are "based on internal characteristics that are difficult to define and assess" (Henry et al., 2013, p. 443). Furthermore, little empirical research has explored the relationship of these dispositional ratings of teacher candidates and their effectiveness as teachers (ibid), though one study of 10 teachers of 80 students focused on the relationship between teachers' locus of control and student achievement gains and found that both male and female students gained more on the achievement measure when their teachers had an internal locus control, compared to gains of students whose teachers have external loci of control (Murray & Staebler, 1974).

Another concern regarding professionalism is that some of the terminology, such as "disposition" and "internal characteristics," suggests that most programs approach a fixed trait rather than an attribute that can be developed. Yet, research on growth mindsets indicates that students perform better when they view intelligence as dynamic rather than fixed. For example, Yaeger and Dweck (2012) find that students who believe that intellectual abilities can be developed tend to show higher achievement following school transitions and have higher completion rates in math courses. These findings have implications not just for how teachers should approach their students, but also for how teacher preparation programs should approach their teacher candidates. Encouraging teacher candidates to adopt a growth mindset can help teachers meet challenges with resilience – which is especially crucial for urban educators working in high-need schools.

As described on the Mindset website, in a growth mindset school: "Teachers collaborate with their colleagues and instructional leaders, rather than shut their classroom doors and fly solo. They strive to strengthen their own practice, rather than blame others."[4] Thus, despite a limited research base that explicitly links measures of professionalism to teacher effectiveness, related work provides theoretical support for the importance of capturing certain behaviors in formative assessments. We focus on evidence that teachers exhibit a growth mindset and behaviors related to teachers' efforts to continually improve their practice. Professionalism in this regard is viewed as "the commitment to continue learning as central to the work of teaching" (Sykes, 1999, p. 245).

UTC Professionalism Measures

UTC as an organization embodies the growth mindset and demonstrates this by providing feedback and support to assist our participants in improving all areas of performance, including professionalism. As such, we use ongoing tracking of behaviors and feedback from site teams and coaches as opposed to measuring teacher dispositions at just one point in time. Our

[4] Accessed online on 10/13/14 from http://www.mindsetworks.com/webnav/whatismindset.aspx.

partner districts, the Baltimore City Schools and the District of Columbia Public Schools, have already adopted professionalism criteria as a part of their teacher evaluation systems. The district measures encompass attendance, punctuality, compliance with district and school policies, respect, and testing integrity.[5] Often, what can be measured reliably (e.g., absences) is really only a small part of the broader construct we wish to capture. UTC gathers data on many of the same elements as our district partners, but adopts a more comprehensive conceptualization of professionalism that is consistent with the inTASC standards and Carol Dweck's notion of a growth mindset.

Six indicators that fall under Strand E (Relentless Pursuit of Continuous Learning) on our TPR:

1) Learner stance: Do participants view teaching as a profession in which we are all working to become more effective and efficient?
2) Continuous learning: Do participants read professionally, apply what is useful, and share their learning with colleagues?
3) Goal setting: Do participants reflect on their practice and set and implement clear goals for improvement?
4) Locus of control: Do participants take responsibility for their successes and failures?
5) Openness: Do participants accept feedback, act on it appropriately, and give respectful feedback to their colleagues?
6) Professionalism: Do participants demonstrate punctuality, respect, and other behaviors of a consummate professional?

In addition, certain indicators are aligned with National Board Certification Proposition 5: Teachers are Members of Learning Communities. Specifically, we seek evidence of our participants' ability to engage proactively with members of their professional learning community.

Coaches provide our participants with feedback on the Strand E indicators in an effort to help them achieve a more positive and productive approach to continuous learning. Coaches rate participants on these six indicators as part of the comprehensive coaching cycles. As such, residents and first year teachers receive formative feedback on professionalism from their coach four times a year, and second year teacher receive formative feedback on professionalism twice a year. By using formative assessments of our participants' professionalism and commitment to continuous learning, we hope to promote a growth mindset among our participants.

Implementation of Professionalism & Growth Mindset as Formative Assessment

In addition to the feedback provided as part of the comprehensive coaching cycles, we found that we needed a way to make our expectations transparent and provide timely feedback in cases where participants are not showing the professional commitment needed to thrive in our program and become effective in the classroom. To this end, we have also instituted a system whereby multiple stakeholders can report back to site teams on participants' professional behaviors on an ongoing basis. In an effort to balance ease of use and comprehensiveness, we have gone through several iterations of our Strand E tracking tool. Our site directors, coaches, coursework instructors, and others who interact with participants can use this online form to report celebrations as well as concerns, so we have timely data on participants' behaviors. Site teams use the data in the tracking tool to acknowledge the accomplishments of individual participants in weekly messages. When a participant has raised multiple concerns, site teams

[5] Baltimore City Public Schools information on professional disposition obtained on 3/25/14 from http://www.baltimorecityschools.org. The District of Columbia Public Schools includes core professionalism as part of IMPACT, the DCPS Effectiveness Assessment System for School-Based Personnel. See http://dcps.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+%28Performance+Assessment%29/An+Overview+of+IMPACT.

may use these data to guide clinical faculty calls and request input regarding what feedback has been provided to the participant and what next steps might be useful.

At the mid-point and end of each school year, site directors fill out a survey for each participant, rating each participant along the dimensions of continuous learning and professional behavior, drawing on data they have gathered in their own interactions as well as the feedback from other constituents via the Strand E tracking tool. We combine the site directors' ratings with the coaches' ratings to generate a combined measure of professionalism. This combined measure constitutes another formative assessment that is shared with participants during mid-year and end-of-year performance reviews.

Ongoing Refinement of Professionalism & Growth Mindset Metrics

While our participants were familiar with Strand E of the TPR, they had less visibility regarding behaviors captured by the Strand E tracking tool. These behaviors are aligned with the TPR, but in an effort to be as transparent with our participants as possible regarding how we track evidence of behaviors, we distributed a copy of the Strand E tracking tool to participants. In addition, the TPR Strand E behaviors as they pertain to coursework are outlined in our course syllabi. In this way, our participants receive consistent messaging regarding the expectations of them as continuous learners from course instructors, coaches, and our site teams.

As a result, since first implementing the online tracking tool just over a year ago, we have had many conversations with site teams regarding what constitutes strong evidence of our participants' professionalism and growth mindsets. We have followed up with site teams when unrelated behaviors appeared in the Strand E tracker. In addition, we streamlined the ongoing tracking tool, eliminating high-inference elements in favor of focusing on concrete, observable actions and added an item for staff to indicate the severity of the behavior observed. We have also developed a rubric for site teams to use when considering the evidence in the tracking tool to create their end-of-semester ratings.

Despite the challenges associated with measuring professionalism and growth mindsets, we see some promising signs in terms of the evidence we gather. Coaches' ratings of our residents as continuous learners improved over the course of SY 2013-14, as seen in Figure 7.

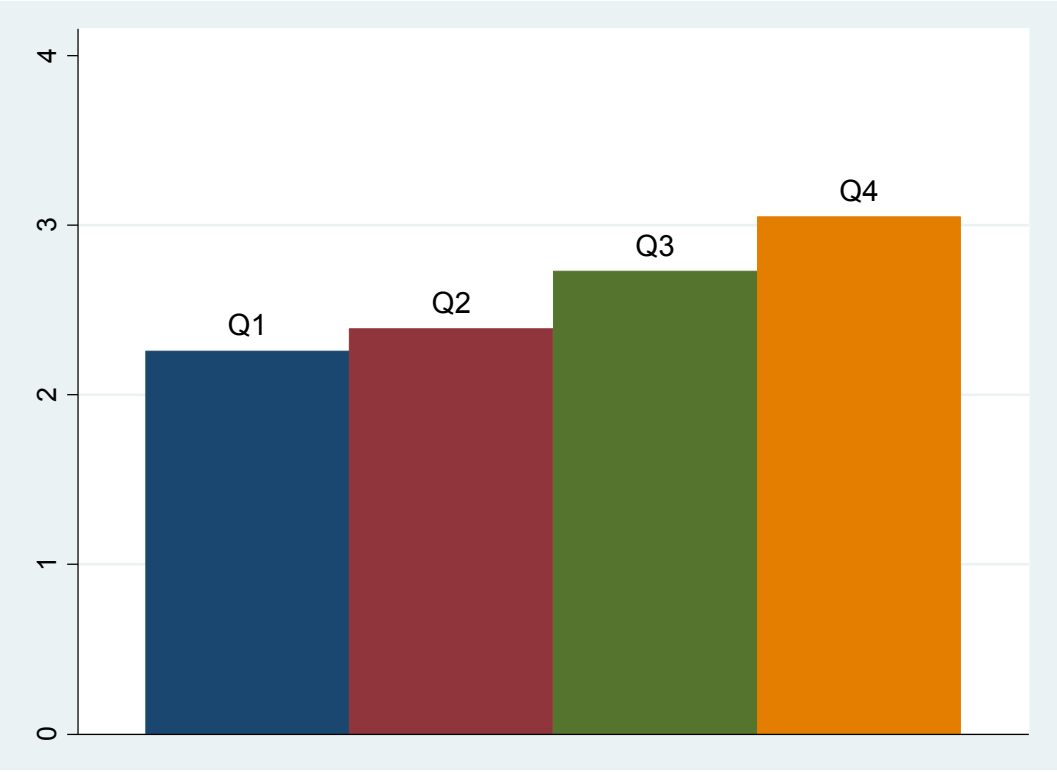

Figure 7. Strand E Ratings of Residents, by Quarter (SY 2013-14).

As we continue to refine how we define and measure professionalism, communication of what we expect from our participants has improved. Clinical faculty report that our residents are more aware of what behaviors are expected of them, and site directors have reported improved behaviors following conversations with participants regarding our expectations.

Challenges with Professionalism & Growth Mindset as Formative Assessment

We face many challenges with adequately capturing participants' professionalism and growth mindsets. As others have noted, it is difficult concept to define and assess such concepts. While Strand E of the TPR provides guidance, we have found that some indicators are not readily observable, such as whether teachers read professionally. Coaches and site directors cannot provide evidence-based ratings of professionalism if the behaviors are not readily observed. In adapting our tracking tools, we have tried to focus on concrete, observable behaviors in an effort to make ratings more objective.

In an analysis of professionalism ratings from SY 2013-14, we found that coaches' assessments of participants were not strongly correlated with site ratings. Because site teams obtain evidence from a variety of sources, including coursework instructors and school partners, we would not expect an exact relationship, but finding an insignificant relationship between the two sets of ratings was unexpected. Coaches and site directors appear to have formed different opinions of participants' professional behavior.

We continue to seek input from clinical faculty and our site teams on the best ways to gather and analyze evidence of professional behavior. We plan to implement focus groups to gather more feedback from stakeholders regarding what evidence they draw on in determining ratings of professionalism and how they interpret various behaviors. The information gathered can be used to inform future training sessions to ensure that all staff share an understanding of what constitutes evidence of professionalism.

FINAL BAR: UTC'S MULTI-MEASURE TEACHER EFFECTIVENESS EVALUATION

UTC's evaluation process culminates in a recommendation for or against certification. Because we are continuously assessing the performance of our participants and providing targeted feedback and coaching support throughout their time in the program, we expect that the majority of the participants who are retained through the end of the third year will meet our standard and be recommended for certification.

We have worked closely with partners from Westat and Education Analytics to refine and strengthen a set of measures that fairly and accurately measure teacher effectiveness for these high-stakes decisions. We combine data on teacher performance across three measures to determine the Teacher Effectiveness Evaluation score. Two of the three measures (classroom practice and professionalism) are also used as formative assessments throughout participants' time in the program; as such, participants receive considerable feedback on their performance in these areas prior to our summative assessment. The three measures are described below.

Classroom practice: During their third year in the program, UTC coaches rate participants on our Teacher Practice Rubric—the same tool used for feedback and accountability purposes from the start of the program—and compute an average score across multiple indicators of teaching practice. This score makes up 40% of the final effectiveness rating.

Student achievement: Students in the classrooms of UTC participants take a computer-adaptive assessment at the start and end of each year. We compare the average growth of students in each classroom to the typical growth rate of students in urban districts to determine student gains. This student gains score makes up 40% of the final effectiveness rating.

Professionalism & Growth Mindset: UTC's Teacher Practice Rubric includes a fifth strand for comprised of indicators of professional behaviors. Similar to the classroom

practice rating, coaches note where participants fall on each of the indicators and compute an average rating across indicators. This coach rating is combined with a survey-based rating from the site director. The combined score makes up 20% of the final effectiveness rating.

The three ratings are then weighted and combined into a composite score (Table 1).

UTC has rolled out this effectiveness standard in phases, and we will use the combined score for the cohort that becomes eligible for certification in June 2015. During the summer of 2014, we reviewed preliminary data to establish a cut score. We will conduct a careful review for any participants who fall below the cut score, taking into account contextual factors that may have influenced scores before making a final decision regarding certification.

	Classroom Practice	**Student Achievement**	**Professionalism & Growth Mindset**
Description of measure	Teacher Practice Rubric score (averaged across strands A-D)	Average student gains for two consecutive years	Ratings on professionalism survey and strand E of the Teacher Practice Rubric
Weight	40%	40%	20%

Table 1: UTC's Effectiveness Standard

Summary

UTC has sought to create a set of multiple, formative assessment measures that capture the skills and practices needed to teach effectively. Drawing on existing research base and external research partnerships provided a useful starting point in developing measures of effective teaching. Prior studies informed our thinking about how to develop formative assessment measures and assess the quality of these measures, and UTC has drawn heavily on this evolving research base to develop sound formative assessments. In the course of implementing UTC's model, we've learned some key lessons regarding the conditions needed to operationalize our vision of high-quality urban teacher preparation. These lessons are described below.

First, we learned that we needed to create a radically explicit connection between coursework and teaching practices across a time continuum, since our residents are simultaneously engaged in coursework and clinical experiences. Coaching cycles are aligned with coursework such that residents are evaluated on the specific teaching practices that they've had an opportunity to practice through key assignments. By first conducting formative assessment of those skills through key assignments in coursework and then observing our participants' implementation of those skills in practice, we hope to scaffold the development of specific teaching skills to support our teachers' growth.

Second, implementing the UTC model with fidelity requires extensive support to our clinical faculty. We've learned that it's not sufficient to provide training at the start of employment. Our Curriculum and Development team provides deep and sustained support to clinical faculty to ensure the quality of coursework and coaching, and to enable consistent use of formative assessment tools. Thus, faculty institutes and monthly meetings are supplemented with faculty mentor calls, coursework materials, paired observations, and observation and feedback from lead clinical faculty. Since our clinical faculty and staff are growing, we need to provide ongoing training and support to ensure that our participants are given ratings that are commensurate with their performance.

Third, having an independent part of the organization monitoring data and holding up a mirror to the organization is instrumental in identifying our own strengths and weaknesses. The Performance and Evaluation team tracks data and provides reports on the progress of our participants. Additionally, because we recognize that inferences about our participants' performance are only as valid as the data that informs them, we monitor implementation of the rubrics and tracking tools we have in place. We examine clinical faculty's completion rates for coaching cycles and expect 100% of comprehensive coaching cycles to be completed on time. In addition, we track data by coach to detect unusual patterns that may indicate more training is needed, such as particularly high or low ratings or lack of differentiation across teachers (rating all teachers, or all aspects of teaching practice, similarly).

Fourth, we recognize that feedback is vital to improvement. Baker et al. (2010) argue that those "seeking to remove ineffective teachers must invest the time and resources in a comprehensive approach to evaluation that incorporates concrete steps for the improvement of teacher performance based on professional standards of instructional practice, and unambiguous evidence for dismissal, if improvements do not occur" (p. 20). In keeping with a view of effective teaching as something that is developed over time, UTC has sought to develop formative assessment measures woven into coursework and clinical practice that are designed to provide teachers feedback on performance as well as targeted coaching support to help them improve.

Feedback also provides the basis for improving programmatic quality. To that end, we have conducted stakeholder surveys and focus groups, and are beginning to implement more systematic reviews of feedback to encourage staff to identify innovative solutions to areas of concern. We conduct annual surveys of clinical faculty to assess whether they are receiving adequate support to implement the coaching and coursework as intended. In spring of 2014, 90% of our clinical faculty reported that the monthly meetings and institutes are helpful. Over 90% agreed that the TPR is a useful instrument for coaching and providing feedback and support to program participants. Regarding adequacy of support, 100% of our clinical faculty reported receiving adequate support to carry out coaching activities and to teach courses; furthermore, all coursework instructors reported having adequate and appropriate materials to teach courses. We've also implemented surveys of stakeholders, including participants, and host teachers, to evaluate the quality of coursework, monitor timeliness of coaching cycles, and obtain feedback on program quality. In spring of 2014, over 90% of our residents and 95% of our first- and second-year teachers agreed or strongly agreed with the following statement: "UTC training gave me the knowledge and skills needed to be effective in the classroom."

Fifth, because we are committed to continuously improving the training and support that we provide our participants, we regularly revise aspects of the program that are in need of improvement. Iterative, robust curriculum work occurs annually. Survey results and course evaluations are used to make adjustments to coursework and to inform programmatic improvements. In addition to revising the curriculum, we plan to continue refining the model based on input from evaluators, as new research emerges in the field, and as we gather more of our own data. We will likely refine our tools as well as our coursework and coaching as we discover answers to these questions. We are also exploring the expansion of our formative assessments to include other measures of teacher effectiveness, such as student surveys.

Finally, ongoing analysis of our measures is essential to continuous improvement. Our external partners, Westat and Education Analytics, have helped us refine our measures and shape them into a multi-year system that we can use confidently to assess the development of every participant. Our partners have conducted a variety of analyses to validate the Teacher Practice Rubric and have found positive correlations between teacher practice ratings and professionalism ratings, which suggests that different facets of effective teaching are related, as we expected. We will continue to analyze the predictive validity of our formative assessments (i.e., the extent to which course grades, the TPR and measures of professionalism are related to student gains) as well as conducting more intensive analyses to understand the effect of various program components.

As we continue to refine and implement a robust new teacher assessment system, we believe UTC's approach can inform teacher educators, policymakers, and education leaders nationwide. Our multi-measure, multi-year teacher evaluation system is one of the first to

measure the performance of early career teachers, tying certification to their effectiveness in the classroom. We hope that others can learn from the system we have developed, as well as the thinking that informs it, and we remain committed to sharing our outcomes and lessons learned as we implement this model with many more new teachers in the coming years.

Acknowledgements

Jennifer Green, Kirsten Mackler, and Roxanne White of Urban Teacher Center made numerous contributions to this paper. They have not only overseen the development and operationalization of UTC's innovative teacher preparation program, but also generously provided guidance and feedback on the analyses described within this paper as well as the structure and content of the paper itself. Celia Parker created the graphics to describe UTC's program and theory of action. We are indebted to our partners at Westat for their analyses of our Teacher Practice Rubric and at Education Analytics for their work developing our student achievement gains model. Both partners provided insightful comments to inform UTC's approach to assessing teacher effectiveness. We also thank Katie Bayerl for assistance in shaping and editing this paper, as well as Lauren K.B. Matlach from the American Institutes for Research and Marisa Goldstein of the Aspen Institute for valuable comments on an earlier draft of this work.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Baker, E.L, Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., and Shepard, L.A. (2010). *Problems with the use of student test scores to evaluate teachers.* Washington, DC: Economic Policy Institute.

Ball, D.L., & Forzani, F.M. (2010). What Does it Take to Make a Teacher? *Phi Delta Kappan, 92*, 8-12.

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Author.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (NBER Working Paper 17699). Cambridge, MA: National Bureau of Economic Research.

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

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Council of Chief State School Officers (CCSSO). (2011). *InTASC model core teaching standards: A resource for state dialogue.* Washington, DC: Interstate Teacher Assessment and Support Consortium.

Council for the Accreditation of Educator Preparation (CAEP). (2013). CAEP Accreditation Standards. Accessed online http://caepnet.files.wordpress.com/2013/09/final_board_approved1.pdf on October 23, 2014.

D'Agostino, J.V. & Powers, S.J. (2009). Predicting Teacher Performance With Test Scores and Grade Point Average: A Meta-Analysis. *American Educational Research Journal, 46*(1), 146-182.

Davis, D.R., Pool, J.E., & Mits-Cash, M. (2000). Issues in implementing a new teacher assessment system in a large urban school district: Results of a qualitative field study. *Journal of Personnel Evaluation in Education*, *14*(4), 285–306.

Davis, E.A., & Boerst, T. (2014). *Designing Elementary Teacher Education to Prepare Well-Started Beginners.* TeachingWorks Working Paper.

Goe, L., Biggers, K., & Croft, A. (2012). *Linking teacher evaluation to professional development: Focusing on improving teaching and learning.* Washington, DC: National Comprehensive Center for Teacher Quality.

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

Greenberg, J., McKee, A., & Walsh, K. (2013). *Teacher prep review: A review of the nation's teacher preparation programs.* Washington, DC: National Council on Teacher Quality.

Guyton, E., & Farokhi, E. (1987). Relationships among academic performance, basic skills, subject matter knowledge, and teaching skills of teacher education graduates. *Journal of Teacher Education, 38*(5), 37-42.

Henry, G.T., Campbell, S.L., Thompson, C.L., Patriarca, L.A., Luterbach, K.J., Lys, D.B., & Covington, V.M. (2013). The Predictive Validity of Measures of Teacher Candidate Programs and Performance: Toward an Evidence-Based Approach to Teacher Education. *Journal of Teacher Education, 64*(5), 439-453.

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

Ho, A.D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill and Melinda Gates Foundation.

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

Kane, T.J., Kerr, K.A., & Pianta, R.C. (2014). *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project.* San Francisco, CA: Jossey-Bass.
Kane, T.J., McCaffrey, D.F., Miller, T., & Staiger, D.O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: Bill & Melinda Gates Foundation.
Kane, T.J., Rockoff, J.E., & Staiger, D.O. (2006). *What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City* (NBER Working Paper 12155). Cambridge, MA: National Bureau of Economic Research.
Kane, T.J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2010). *Identifying effective classroom practices using student achievement data* (NBER Working Paper 15803). Cambridge, MA: National Bureau of Economic Research.
Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality.
Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.
Murray, H.B., & Staebler, B.K. (1974). Teacher's locus of control and student achievement gains. *Journal of School Psychology, 12*(4), 305-309.
National Board for Professional Teaching Standards (NBPTS). (2014). http://www.nbpts.org/members-learning-communities
Putman, H., Greenberg, J., & Walsh, K. (2014). Training our future teachers: Easy A's and what's behind them. Washington, DC: National Council on Teacher Quality.
Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.
Rockoff, J.E., Jacob, B., Kane, T. J., & Staiger, D. O. (2008). *Can you recognize an effective teacher when you recruit one?* (NBER Working Paper 14485). Cambridge, MA: National Bureau of Economic Research.
Rockoff, J.E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.
Rockoff, J.E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review: Papers and Proceedings 100*: 261–266.
Rowan, B., Correnti, R., & Miller, R.J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *Teachers College Record*, 104(8), 1525–1567.
Sanders, W.L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.
Staiger, D., & Rockoff, J. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives, 24*(3), 97–118.
Sykes, G. (1999). The "New Professionalism" in Education: An Appraisal. In Murphy, J., & Seashore Louis, K. (Eds.). Handbook of Research on Educational Administration (2nd Edition). San Francisco, CA: Jossey-Bass.
Taylor, E.S., & Tyler, J.H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers* (NBER Working Paper 16877). Cambridge, MA: National Bureau of Economic Research.
TNTP, & Student Achievement Partners. (2013). *Fixing classroom observations: How the Common Core will change the way we look at teaching.* New York, NY: Author.
Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89–122.
Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: TNTP.
Wilson, B. & Robinson, V. (2012). Predicting Teaching Performance: Proceed with Caution. *Journal of Assessment and Accountability in Educator Preparation, 2*(1), 58-61.

Yaeger, D.S. & Dweck, C.S. (2012). Mindsets That Promote Resilience: When Students Believe That Personal Characteristics Can Be Developed. *Educational Psychologist, 47*(4), 302–314, 2012.