ASSESSING EFFECTIVENESS:
HOW URBAN TEACHERS EVALUATES ITS NEW TEACHERS

2016

Cara Jackson, PhD and Kirsten Mackler
# TABLE OF CONTENTS

**EXECUTIVE SUMMARY**  
1

**PART I: BACKGROUND**  
2
The Need for New Measures  
2  
Developing Great Teachers  
3  
Learning as We Grow  
4

**PART II: OVERVIEW OF THE TEACHER PERFORMANCE ASSESSMENT SYSTEM**  
6
Years 1-3: High Expectations at Every Stage  
6  
Final Bar: A Multi-Measure Teacher Effectiveness Evaluation  
7

**PART III: URBAN TEACHERS' EFFECTIVENESS EVALUATION IN DETAIL**  
9
Rationale for a Multi-Measure Standard  
9  
Measure 1: Classroom Observation Ratings  
10  
Urban Teachers' Rubric Design  
10  
Rubrics as Powerful Diagnostic Tools  
11  
Limitations to Observation Ratings  
11  
Assessing Rubric Validity  
12  
Measure 2: Student Gains  
15  
Urban Teachers' Student Gains Model  
15  
Limitations of Student Achievement-Based Measures  
16  
Measure 3: Growth Mindset and Professionalism  
18  
Ongoing Refinement  
19

**ACKNOWLEDGEMENTS**  
20

**REFERENCES**  
21
EXECUTIVE SUMMARY

Urban Teachers was founded with a simple yet ambitious goal: to provide high-need schools with effective new teachers. Since 2010, we have worked to enact that vision, building an innovative, research-informed teacher preparation program from the ground up. A hallmark of this work—and the result of much careful consideration—is Urban Teachers’ Teacher Performance Assessment System. The Teacher Performance Assessment System is a multi-year and multi-measure system that helps us track the progress of every new teacher in our program, determine who will advance from one stage to the next, and make our ultimate evaluation of effectiveness.

Urban Teachers is one of a very small number of teacher preparation programs to take the step of developing a performance assessment system. In the absence of a broader accountability system for teacher education, we have invested in creating our own measures because we believe that, despite the challenges involved, teacher preparation programs play a crucial gatekeeping role for schools and that we must shoulder the responsibility of verifying that new teachers are equipped for the job.

Our approach has evolved over time. In the past two years, we engaged with external partners at Westat and Education Analytics to validate our measures and combine them into a single standard of effectiveness, which we use to determine which participants to recommend for full teacher certification. This paper describes the results of that collaboration, the core elements of our system, and the decisions that have led us to the current design.

As one of the first teacher preparation programs in the country to implement a robust evaluation system for our participants, we believe this approach and the lessons we have learned during the design process can inform teacher educators, policymakers, and education leaders nationwide.
PART I: BACKGROUND

THE NEED FOR NEW MEASURES

Extensive research has demonstrated that, of all in-school factors, it is the quality of the interactions between teachers and students that has the greatest impact on student learning. Evidence from a large-scale study in which teachers were randomly assigned to classrooms indicates that teachers identified as more effective produced greater student achievement growth in the following school year relative to other teachers in the same school, grade, and subject (Kane, McCaffrey, Miller, & Staiger, 2013). Students assigned to high value-added teachers are more likely to attend college, attend higher-ranked colleges, earn higher salaries, live in higher socioeconomic status neighborhoods, and save more for retirement (Chetty, Friedman, & Rockoff, 2014).

Given the significant and far-reaching impacts of quality teaching, many have advocated for including measures of teacher effectiveness in policies that govern decisions about teacher hiring, certification, tenure, and advancement (Glazerman, Loeb, Goldhaber, Staiger & Raudenbush, 2010; Gordon, Kane & Staiger, 2006). They argue that, without information about how teachers perform, schools and districts cannot make strategic choices about staffing their schools, supporting their students, or determining whom to promote into leadership positions. Recent evidence from Washington, DC's teacher evaluation system lends credence to this position, as results suggest that inducing lower-performing teachers to exit the system can have positive impacts on student achievement (Adnot, Dee, Katz, & Wyckoff, 2016).

Though more states and districts have begun using effectiveness measures to inform strategic decisions regarding teachers who are already in the classroom, pre-screening of new teachers remains challenging.¹ While qualities like academic credentials are often given weight in the hiring process, research finds the relationship between degrees or coursework and student achievement to be inconsistent and often statistically insignificant (c.f Aaronson, Barrow, & Sander, 2007; Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2006; Clotfelter, Ladd, & Vigdor, 2007). Currently, only about a quarter of traditional teacher preparation programs routinely gather information on the performance of their teacher candidates (Greenberg, McKee & Walsh, 2013). That leaves schools with little information to go on when deciding whom to hire, making each decision more or less a gamble. This lack of information is particularly problematic in high-need urban schools, where a single year of quality instruction can make an enormous difference to students but where a

¹ Using data from Chicago and Texas, respectively, Aaronson et al. (2007) and Rivkin et al. (2005) both find that observable teacher characteristics explain very little of the variation in teacher effectiveness. Using data from North Carolina, Clotfelter et al. (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.
revolving door of underprepared new teachers leave too many children with teachers unequipped to meet their needs.

DEVELOPING GREAT TEACHERS

Urban Teachers seeks to provide high-need schools with effective teachers through a combination of intensive, multi-year support and ongoing accountability.

First, we require every participant to commit to working with us for four years. They spend the initial fourteen months, the residency year, working in classrooms with guidance from host teachers and on-site coaching from our expert faculty. At the same time, they take graduate-level courses that introduce them to best practices in the field and provide immediate opportunities to try those practices with students.

Participants spend the next three years working as full-time teachers of record in Urban Teachers’ partner schools, where they continue to receive regular, on-site coaching and support. At the end of the first year as teachers of record, successful participants earn a master’s degree, and at the end of the following year (their third year with Urban Teachers), they become eligible for dual certification in their content area and special education. Along the way, our participants undergo ongoing evaluation of teaching practice, growth mindset, and professionalism. Figure 1 shows how Urban Teachers ensures teacher effectiveness through rigorous selection, intensive training and support, and ongoing evaluation, resulting in more effective teachers for our school partners.

The delayed decision around certification is key; it is what allows Urban Teachers 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 gathered multiple sources of data that demonstrates their ability to implement best practices and improve student achievement in real classrooms.
LEARNING AS WE GROW

When Urban Teachers began this work in 2010, new teacher evaluation was still a nascent field. A few teacher preparation organizations were exploring methods of assessing their teacher candidates, but no single, reliable approach had yet emerged. We drew on the research that was available at the time, as well as our own experiences as educators and education leaders.

Our original concept was straightforward (Urban Teacher Center, 2009). We would employ two measures of effectiveness:

- **Grade point average (GPA):** With coursework designed around best practices in the field and embedded in the classroom experience, we would use participants’ course grades as a proxy for capturing their emerging competency in the classroom.

- **Student achievement gains:** Participants would be responsible for showing an average of at least one year of student achievement gains for each of their first two years as teachers of record; as a first step in that direction, residents would be responsible for showing significant gains with a small group of students during their five-week practicum.

By the second year of implementation, we recognized that while we were headed in the right direction, our model needed revision. In particular, we recognized that:

- **We needed timely and actionable indicators of performance:** Early on, we realized that our coaches were investing considerable time with participants who were not on track to becoming effective. Many of these participants eventually elected to leave the program but not before draining valuable time and resources. We realized that course grades were not enough to monitor performance and that our coaches and site directors were noticing other, observable behaviors that could more quickly and accurately inform our decisions about program continuation.

- **Student achievement gains were not a useful measure for the residency year:** Urban Teachers residents take responsibility for a small group of students during their official practicum. We found that student data we gathered from these groups were highly variable due to small sample size and that it was impossible to separate the effect of the resident from the effect of the host teacher. Ultimately, we decided student achievement gains were not a fair measure of our residents’ performance, but we remained committed to using student achievement gains during the teaching years.

- **We needed better assessments of student achievement:** The first assessments we tried, a Fountas & Pinnell reading assessment and a Stanford mathematics assessment, could not easily be converted into a growth measure. We tried using the Stanford 10 next, but the paper-and-pencil tests proved cumbersome to administer and costly. We reviewed other
available assessments and selected NWEA’s Measures of Academic Progress (MAP) assessments; these nationally normed, computer-adaptive tests are easier to administer, more motivating for students, and less vulnerable to compromise or error.

- **We needed to consider multiple measures for our final determination of effectiveness:**
  Initially, we had intended to use student achievement gains as our final assessment in year three, but we recognized that every measure had some degree of error and that, as we began to collect observational data in classrooms, we had other indicators of effectiveness that, when combined, offered a more complete picture of our participants’ performance.

By end of our second year, we had begun to shift our approach to evaluating effectiveness, but we needed help formalizing our assessment system. We enlisted support from Westat and Education Analytics to review our proposed model and make recommendations about how to use the measures we had selected to determine effectiveness. They conducted a thorough review of related research, tested the validity of our measures with preliminary data, and recommended a straightforward method for applying these measures uniformly to all future participants.

We introduce the pieces of the resulting approach—Urban Teachers’ Teacher Performance Assessment System—in the sections that follow. First, we describe how we assess the progress of participants during their first three years in the program, making decisions about who will advance through the program and who will not. The remainder of the paper focuses on the multi-measure Teacher Effectiveness Evaluation that will guide our most important decision: whom we recommend for certification.
PART II: OVERVIEW OF THE TEACHER PERFORMANCE ASSESSMENT SYSTEM

Urban Teachers’ Teacher Performance Assessment System (T-PAS) encompasses more than a single assessment of effectiveness. We gather and use data over a three-year period, which leads to our final determination of effectiveness. Here, we describe what that process looks like over time.

YEARS 1-3: HIGH EXPECTATIONS AT EVERY STAGE

Urban Teachers gathers data on participants over three years prior to making a decision about certification. These data guide the support that coaches provide, inform our investments as a program, and inform our deliberate approach to attrition. During each year of the program, teacher candidates must meet expectations across four domains to remain in good standing.

1. **Course grades:** Participants must maintain a 3.0 grade point average to remain in good standing. Residents who fail a course are dismissed immediately, pending review. First-year teachers, who take a lighter course load, are held to the same grading standard, but dismissal decisions are not made until the end of the school year to minimize disruption to students. First-year teachers may also be allowed to retake the course.

2. **Teaching practice:** Participants are rated on the Teacher Practice Rubric (TPR) by multiple observers and on multiple occasions. These detailed, evaluative classroom observations are in addition to non-evaluative visits that are part of Urban Teachers’ ongoing coaching on the teaching practices delineated in the TPR. Participants who do not meet the minimum bar (which increases over time) meet with their site teams and instructional coach to discuss next steps. If they continue to fail to meet Urban Teachers expectations, they are eligible for dismissal.

3. **Growth mindset & professionalism:** One strand of the Teacher Practice Rubric is devoted to the mindset and behaviors associated with an effective teaching career, such as being open to feedback and pursuing continued improvement in practice. Coaches and site teams gather data on these indicators for all participants. Participants who do not meet expectations receive feedback from their coaches and site team. If they continue to fail to meet Urban Teachers expectations, they are eligible for dismissal.

4. **State and local requirements:** To earn state licensure, participants must pass state-administered licensure exams in the subject they teach. The assessments vary by state and by subject; participants typically take a set of exams that measure academic skills in reading, writing, and mathematics prior to enrollment, exams that measure general and subject-
specific teaching skills and knowledge during their residency year, and exams that measure
general pedagogical knowledge prior to certification. Urban Teachers' participants must also
remain in good standing with their school and district; Urban Teachers cannot guarantee
continued participation for teachers whose contracts have not been renewed.

Urban Teachers provides support to ensure that as few participants as possible fall below these
expectations. We provide participants with performance reports that clearly indicate their ratings on
all indicators in comparison with program expectations. In addition, coaches and site directors will
often intervene when they observe a teacher or resident fall behind, providing feedback, action
steps, and support to get them on track before a dismissal conversation is required.

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 investing our energies in those who are. Each year, approximately 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.

As a result of this multi-year assessment process, by the time our participants become eligible for
certification, we can focus on those who have already met previous benchmarks and have shown
true commitment to urban teaching.

**FINAL BAR: A MULTI-MEASURE TEACHER EFFECTIVENESS EVALUATION**

Urban Teachers' evaluation process culminates in a recommendation for or against certification. We
have worked closely with partners from Westat and Education Analytics to refine and strengthen a
set of measures that fairly assess teacher effectiveness for this high-stakes decision. We plan to only
recommend teachers for certification who meet high expectations for classroom performance
across three measures:

**Teaching practice:** During their third year in the program (and second year as teachers of record),
participants are rated by their coach and by an independent observer on our Teacher Practice
Rubric—the same tool used for feedback and accountability purposes from the start of the program.
We compute an average score across four areas of practice. This score makes up 40% of the final
effectiveness rating (or 70%, for those who do not have student gains data).

**Student gains:** Students in the classrooms of Urban Teachers 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. This student gains score
makes up 40% of the final effectiveness rating.
**Growth mindset & professionalism:** the Teacher Practice Rubric includes a fifth strand comprised of indicators of professional behaviors and attitudes. Similar to the teaching practice rating, coaches and site teams note where participants fall on each of the indicators and compute an average rating across indicators. The combined score from their third year in the program makes up 20% of the final effectiveness rating (or 30%, for those who do not have student gains data).

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

**TABLE 1: URBAN TEACHERS’ TEACHER EFFECTIVENESS EVALUATION**

<table>
<thead>
<tr>
<th>Description of measure</th>
<th>Teaching Practice Rubric score (averaged across strands A-D)</th>
<th>Average student achievement gains for two consecutive years</th>
<th>Average ratings on indicators in strand E of the Teacher Practice Rubric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight with student gains</td>
<td>40%</td>
<td>40%</td>
<td>20%</td>
</tr>
<tr>
<td>Weight without student gains²</td>
<td>70%</td>
<td></td>
<td>30%</td>
</tr>
</tbody>
</table>

Urban Teachers has rolled out this effectiveness evaluation in phases. During the summer of 2014, we reviewed preliminary data and established a rigorous but fair cut score. Going forward, a panel of reviewers, including staff familiar with participants and their school environments, will conduct a careful review of any participants who fall below the cut score, taking into account contextual factors that may have influenced ratings and the participants’ performance over time before making a final decision regarding certification.

² We do not collect student gains data for pre-kindergarten teachers or teachers of fewer than 10 students. In addition, due to concerns about misattributing gains, we do not collect student gains for participants who are co-teachers or are in push-in or pull-out positions.
PART III: URBAN TEACHERS’ EFFECTIVENESS EVALUATION IN DETAIL

The final section of this report focuses on three measures we use to inform our decision about whom to recommend for full certification. As described above, our Teacher Effectiveness Evaluation takes place at the end of participants’ third year with Urban Teachers. Here, we offer more detail about the research and thinking behind each measure, as well as the adjustments we have made—and continue to make—to ensure fair and consistent application of our effectiveness standard.

RATIONALE FOR A MULTI-MEASURE STANDARD

Teaching is a complex, multifaceted job, and any single measure of performance reflects only a limited part of the larger construct of teaching quality (Bell, Little, Croft & Gitomer, 2009). An effectiveness evaluation that encompasses multiple measures has the potential to capture a greater range of the many actions and decisions that underlie effective teaching and thus generate more valid inferences about overall effectiveness.

A growing body of research indicates that combined teacher performance measures are more reliable and have greater power to predict future student achievement than any single measure. For example, principals’ assessments of teachers and student growth measures, when used in combination, may be a better predictor of future student achievement than either measure alone (Jacob & Lefgren, 2008). Similarly, researchers affiliated with the Measures of Effective Teaching (MET) Project have found that individual measures each have relative strengths and limitations: student growth measures tend to be the most powerful predictors of future student performance, but they are less reliable than data from student surveys and data derived from multiple classroom observations. MET Project researchers conclude that a composite measure that includes all three forms of data is the strongest and most reliable predictor of future student achievement gains.3

Urban Teachers’ effectiveness standard combines three measures into a single, summative score, applying greater weight to classroom observation ratings and student gains data than to growth mindset and professionalism ratings. Our approach is in keeping with research that suggests that outcomes-based measures (such as estimates of teachers’ impact on student learning) and

---

3 Mihaly et al. (2013) note that the ability of composite measures to predict future performance depends on the outcome measured. Composites with relatively equal weight on each measure will tend to capture the component of effective teaching that is common across measures and will be a reasonably good predictor of future performance on any given measure. Composites that more heavily weight a particular measure will be a better predictor of future teacher performance on that particular measure but a worse predictor of future performance on other measures.
performance-based measures (such as data from classroom observations) both predict the achievement of teachers’ future students, and that each type of measure contains information distinct from the other (Rockoff & Speroni, 2010).

Our scoring process is compensatory. In theory, a high rating in one area can make up for a lower rating in another. In practice, selective attrition should have already removed those candidates who were not meeting expectations for classroom practice and growth mindset and professionalism. By the time participants are eligible for certification, it would be unusual to find strikingly divergent ratings across the three measures. Such divergence would trigger a careful review of data and related documentation.

**MEASURE 1: CLASSROOM OBSERVATION RATINGS**

Teacher practice rubrics are intended to offer fair, reliable judgments of teaching quality based on a shared understanding of what an effective learning environment looks like. 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 gains (Kane, Taylor, Tyler & Wooten, 2010). 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.

**URBAN TEACHERS’ RUBRIC DESIGN**

The Teacher Practice Rubric (TPR) is aligned with the standards set by the Council of Chief State School Officers’ Interstate Teacher Assessment and Support Consortium (InTASC) and encompasses four sets of skills that new teachers must master in order to become effective. Urban Teachers’ coursework and coaching support is built around these four areas of practice:

- **Strand A**: Build a productive and nurturing classroom environment.
- **Strand B**: Operate as a diagnostician, 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. Observers assign ratings for each indicator on a four-point scale; the final rating is the average across all indicators.
RUBRICS AS POWERFUL DIAGNOSTIC TOOLS

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. Observation of classroom practice by experienced educators, guided by a rubric that describes strong practice, facilitates valuable conversations regarding novice teachers’ work. Empirical evidence suggests that such conversations can support improvements in instructional practice. In this regard, teacher practice rubrics are not merely a mechanism for evaluating effectiveness but can actually promote effectiveness.

Teachers and teacher candidates need and deserve such diagnostic support before rubrics are used for high-stakes decisions. In keeping with recommendations from researchers, we have developed a comprehensive approach to evaluation that incorporates steps for improvement based on the TPR (Baker et al., 2010). As participants advance through the program, they receive feedback on their performance in the classroom that is based on our rubric. Coaches identify areas of strength and weakness, providing strategies for improvement and one-on-one coaching over three years. We gather evidence from multiple sources, including staff who are familiar with the participants and their school settings, when a participant is at risk of dismissal.

LIMITATIONS TO OBSERVATION RATINGS

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). This lack of variation in observational ratings is problematic in light of research that suggests teachers vary widely in their ability to produce student achievement gains (Aaronson, Barrow & Sander, 2007). 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

---

4 For example, Taylor and Tyler (2012) 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.
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.⁵

Many of the limitations of observation rubrics can be mitigated by investing in training observers and by conducting appropriate oversight to ensure fidelity of a rubric's use.⁶ 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. In general, adding raters improves reliability more than increasing the number of observations by the same rater (Hill, Charalambous & Kraft, 2012).

Since Urban Teachers uses data from the observation rubric as the basis for consequential decisions, it is important that our observers apply the rubric uniformly. We train and calibrate observers before they begin classroom observations, and we have invested in an online platform to improve consistency in how observers record their ratings. We have also created norming protocols to ensure coaches are using the tool consistently. In keeping with the findings from the MET Project and our own work suggesting that multiple observations increase data reliability, in school year 2015-2016 we added independent observers to the evaluation process so that we can draw on ratings from multiple observers and multiple observations for high-stakes decisions (Ho & Kane, 2013; Jia, Cummings, Jackson, Clifford, & Hoch, 2015).

**ASSESSING RUBRIC VALIDITY**

Urban Teachers, 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. 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 cohort 2010 and r=0.34 for cohort 2011).⁷ Urban Teachers confirmed these results with subsequent cohorts, finding that that the average of a participant's TPR ratings across a school year is significantly related to student gains (Jackson, 2015). Although these analyses are based on a relatively small sample of

---

⁵ On the various instruments tested by Measures of Effective Teaching project, a single observation by a single rater generated reliability in the 0.14 to 0.37 range. By comparison, researchers typically report reliability of student growth measures in the 0.30 to 0.50 range.

⁶ For example, raters participating in the Measures of Effective Teaching 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.

⁷ Our external partner, 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 participants, the population parameters were not estimated, but instead calibrated from a large sample from urban districts and then assumed to be accurate for students and teachers associated with Urban Teachers.
Urban Teachers 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 begin their careers with a range of prior experiences and whose skills may develop at different rates. Urban Teachers’ observational ratings do show substantial variation and do not cluster strongly in the middle or top categories (Figure 2). There is no indication of the leniency that TNTP reported finding in other teacher performance evaluation systems (Weisberg et al., 2009).

**Ratings improve as teachers develop their practice.** Westat’s longitudinal analysis of TPR ratings finds that ratings of Urban Teachers’ 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. Figure 3 displays the ratings of three cohorts during the course of school year 2014-2015. 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 quarter are lower than average ratings for the fourth quarter. Second, experienced teachers perform better on average than novice teachers; averages for the fourth quarter indicate that cohort 2012 (in their second year as teachers) ended the school year with the highest ratings on average, while cohort 2014 (candidates still in their residency year) had the lowest ratings on average.

---

8 On average, measures of teachers’ impact on student achievement gains increase in the first few years of teaching. Several studies have documented a leveling off of returns to experience. For example, Rivkin et al. (2005) find that first-, second- and third-year teachers perform worse than their more experienced counterparts, but any gains beyond the first three years are small and not statistically significant in both mathematics and reading. Clotfelter, Ladd, and Vigdor (2006) and Clotfelter et al. (2007) find that about half the returns to experience occur within first one or two years of experience.
FIGURE 2: DISTRIBUTION OF STRAND AVERAGE RATINGS, SY 2014-2015, YEAR 1 TEACHERS

FIGURE 3: TEACHER PRACTICE RATINGS BY YEAR IN PROGRAM, SY 2014-2015
MEASURE 2: STUDENT GAINS

Numerous studies over the past decade have demonstrated there is considerable variation in the extent to which individual teachers influence student achievement (Aaronson et al., 2007; Atteberry, Loeb, & Wyckoff, 2015). Statistical models estimate the impact of a teacher on student performance by comparing that students’ test scores from one point in time to another, while controlling for factors (such as family income) that are beyond that teacher’s control. Longitudinal studies have demonstrated a link between the value-added estimates of teachers early in their careers and their performance in subsequent years; on average, teachers in the lowest quintile of effectiveness early in their career have significantly less impact on student learning than teachers in other quintiles even five years later (Atteberry, Loeb & Wyckoff, 2015; Goldhaber & Hansen, 2010). Early-career value-added estimates predict far more of the variation in teachers’ future effectiveness than other measures, such as SAT scores or the competitiveness of their undergraduate institutions (Atteberry et al., 2015).

URBAN TEACHERS’ STUDENT GAINS MODEL

Given the power of student growth measures to predict future performance, it makes sense to include a measure of teachers’ impact on student learning in early-career evaluation—and in Urban Teachers’ case, prior to certification. Urban Teachers was able to learn from the challenges of other teacher certification programs that lacked quality student learning data to include as a measure of teacher effectiveness. We started small, launching our program in a limited set of subject areas for which reliable assessments were available. Instead of depending on state or district data, we decided to invest in implementing pre- and post-assessments in participating classrooms so we would have a consistent measure of student achievement gains.

After some trial and error, described in Part II, we selected NWEA’s computer-adaptive Measures of Academic Progress (MAP) assessments, which adjust the difficulty of questions as students respond, resulting in more precise estimates of student skills and understanding. Some of our partner schools and districts implement the assessment themselves, and in those cases, we request access to the data. In the remaining classrooms, we implement the assessments ourselves in the spring and fall. Students take either a reading or math assessment, depending on the subject specialty of their teacher.

Initially, our intention was to look for an average of at least a year of growth in every teacher’s classroom, but because the field has not yet converged on a single definition of what constitutes a year of growth, we took the advice of our external partners and decided to use a common proxy: the average growth of students in a normed sample. We compare our teachers’ classrooms to the average growth of students in the same grade level and subject from a sample of large urban districts.
We begin with an estimate of the average student gains for each teacher. These estimates are in standard deviation units, falling along a bell curve with zero at the center. A teacher whose students, on average, have similar gains to other urban students receives an estimate of zero. Teachers whose student gains fall below the average receive a negative score, and those with student gains that exceed the typical growth rate receive a positive score. There is some uncertainty in these estimates, so for each estimate we also obtain the standard errors, which can be used to calculate the confidence interval. The confidence interval represents a range of scores; there is a 95 percent chance that a teachers’ true student gains is within this range and the true value of student gains could plausibly lie anywhere in that range. Researchers have recommended accounting for the confidence interval surrounding estimates of student gains (ASA, 2014; Di Carlo, 2012). We use the confidence interval to categorize our teachers, designating a score from 1 to 4, with 3 representing average student gains.

**TABLE 2: STUDENT GAINS CATEGORIZATION**

<table>
<thead>
<tr>
<th>Category</th>
<th>Performance</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Low</td>
<td>The upper bound of the confidence interval falls at least 1 standard deviation below the average (which is zero). Student gains are significantly far below the gains of a typical first-year teacher.</td>
</tr>
<tr>
<td>2</td>
<td>Below Average</td>
<td>The upper bound of the confidence interval is less than 1 standard deviation below the average and does not cross zero. Student gains are significantly below the gains of a typical first-year teacher.</td>
</tr>
<tr>
<td>3</td>
<td>Average</td>
<td>The confidence interval crosses zero. Student gains are not significantly different than the gains of a typical first-year teacher.</td>
</tr>
<tr>
<td>4</td>
<td>Above Average</td>
<td>The lower bound of the confidence interval is above zero and does not cross zero. Student gains are significantly higher than the gains of a typical first-year teacher.</td>
</tr>
</tbody>
</table>

**LIMITATIONS OF STUDENT ACHIEVEMENT-BASED MEASURES**

Student achievement-based measures of teacher effectiveness are open to uncertainty. They provide an estimate of a teacher’s effect, but there is no way to ascertain to what extent the estimate reflects the real teacher effect. An individual teacher's gains score might vary from year to year or across groups of students. Researchers have found only moderate year-to-year correlations, ranging from 0.2 to 0.6 (Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood & Mihaly, 2009). While this variability may reflect changes in teachers’ true effectiveness from one year to the next, when
making high-stakes decisions it is crucial to take steps to address and account for uncertainty to the extent possible.

Urban Teachers’ model makes adjustments to account for sampling error and other factors that might cause uncertainty in growth estimates. First, we only consider student achievement gains for those teachers linked to ten or more students, as variability due to sampling error decreases as class size increases. In addition, we use a Bayesian shrinkage adjustment, which shrinks less reliable estimates towards the mean. That is, our estimates implicitly assume that a teacher is average unless strong evidence shows otherwise (Harris, 2011). Finally, in keeping with research that indicates that additional years of data improve the predictions of future effectiveness for early career teachers (Atteberry, Loeb, & Wyckoff, 2015), Urban Teachers averages participants’ student gains estimates across the first two years in which our participants are teachers of record. Averaging gains across multiple years substantially reduces the variation in estimated performance. In an effort to account for differences between urban and non-urban districts, our student gains are estimated relative to student gains in urban districts.

Researchers have pointed out several practical concerns with evaluating teachers based on student test scores, including availability of appropriate tests and summer learning loss (Baker et al., 2010; Gershenson & Hayes, 2016). To the extent possible, Urban Teachers has taken steps to address these concerns. Students of Urban Teachers’ first- and second-year teachers are administered vertically aligned, computer-adaptive assessments, which allows us to capture student growth, even for students who do not begin on grade level. Because assessments are administered in fall and spring, differential summer learning does not produce bias.

A final issue with evaluating teachers based on student test scores has to do with problems of attribution; given the collaborative nature of teaching within schools, multiple educators may have an influence on student learning in any given year (Valli, Croninger & Walters, 2007). For that reason, we do not use student gains data for our residents who share responsibility for instruction with host teachers and will not make evaluative decisions using student data for teacher specialists who “push in” to existing classrooms or “pull out” students for small-group and individual instruction, as their students’ learning will be influenced by other teachers. Altogether, we are able to collect student

---

9 We expect that our teachers are improving over time. In the small number of cases where a teacher is close to the cutoff for being recommended for certification, we may re-run the evaluation model using the student gains estimate from the second year of teaching.

10 Goldhaber and Hansen (2013) note that more years of data on each teacher would improve the reliability of value-added estimates, but the marginal improvements in predictive power are highest when moving to the two- or three-year measures, and improve little with four or more years of data.

11 Since student gains measures are normative, teachers are essentially compared to one another. To avoid creating perverse incentives and undermining collaboration among teachers, Urban Teachers does not compare our participants to one another.
gains data for almost half of our participants. As a comparison, districts typically have such data only for teachers in tested subjects and grade levels. In the District of Columbia, this is 15% of the teaching population (Jacob, Rockoff, Taylor, Lindy, & Rosen, 2016).

MEASURE 3: GROWTH MINDSET AND PROFESSIONALISM

Urban Teachers’ third and final measure of effectiveness captures our participants’ attitudes and behaviors as professionals. We realized early on that we needed a way to make our expectations transparent and act swiftly in cases where participants were not showing the professional commitment needed to thrive in our program and become effective in the classroom.

Two of our partner districts, the Baltimore City Public Schools and the District of Columbia Public Schools, have incorporated professionalism criteria into their teacher evaluation systems, measures that encompass attendance, punctuality, compliance with district and school policies, respect, and testing integrity. Urban Teachers gathers data on many of the same elements, but we have adopted a broader conceptualization of professionalism in keeping with recommendations from the Council of Chief State School Officers’ InTASC. We assess participants across four domains of growth mindset and professionalism:

1. Professional learning community: Do participants engage with peers by listening, posing questions, and demonstrating self-awareness?

2. Openness to feedback: Do participants seek out feedback, respond to feedback in a professional manner, and implement the feedback they receive?

3. Locus of control: Do participants take responsibility for continually improving their teaching practice?

4. Professionalism: Do participants demonstrate punctuality and engage in a respectful manner with staff, peers, school administration, and students?

We ask instructional coaches, coursework instructors, and Urban Teachers staff who interact with participants to document incidents of concern and those worthy of celebration, so we have a record of participants’ behaviors to determine trends. Urban Teachers staff draw on data gathered based on their own interactions with the participant as well as conversations with other constituents, including school administrators. Participants are provided feedback and are rated along these dimensions on multiple occasions throughout the year to assess progress.

---

Baltimore City Public Schools information on professional disposition obtained on March 25, 2014 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://www.dc.gov
ONGOING REFINEMENT

Urban Teachers has drawn heavily on an evolving research base to develop a sound teacher evaluation system, one that is designed not just for accountability purposes but also to help teachers improve. Over time, we have discovered which measures allow us to capture the elements of effective teaching practice and learned how to operationalize our vision of a standard of effectiveness.

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 support and assess the development of every participant. Our partners have conducted a variety of analyses to validate the core components of our evaluation model and have found positive correlations between teacher practice ratings, student gains, and ratings on growth mindset and professionalism.

Our work is not done. We will continue refining the model based on input from stakeholders, new research emerging in the field, and further analyses on our data. We recognize that our inferences about teacher effectiveness are only as valid as the underlying measures. Through our ongoing research agenda, we will continue to assess the predictive validity of each measure as well as the composite standard of teacher effectiveness. We are particularly interested to learn to what extent observation rubric ratings predict student gains from one year to the next and which classroom practice indicators are most strongly associated with student gains. We will likely refine our tools and processes as well as our coursework and coaching as we discover answers to these questions. We are also exploring the expansion of our evaluation model to include other measures of teacher effectiveness, such as student surveys.

While our learning continues, the work we've done so far has helped us develop a more comprehensive, valid, and reliable approach to evaluating new teachers. Our multi-measure, multi-year Teacher Performance Assessment System is one of the first to assess 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

Urban Teachers is grateful to our partners at Westat for their analyses of our Teacher Practice Rubric and Education Analytics for assisting us in calculating student achievement gains. Both partners provided insightful comments to inform Urban Teachers’ approach to assessing teacher effectiveness. Celia Parker of Urban Teachers provided invaluable assistance with graphics. In addition, we received excellent feedback from reviewers Lauren K.B. Matlach from the Rhode Island Department of Education and Marisa Goldstein of the Aspen Institute. Finally, we are deeply indebted to Kim Fleming and Katie Bayerl for assistance in shaping and editing this paper.
REFERENCES


TNTP & Student Achievement Partners. (2013). *Fixing classroom observations: How the Common Core will change the way we look at teaching.* New York, NY: TNTP.


CONTACT INFORMATION

CARA JACKSON
ASSISTANT DIRECTOR OF RESEARCH & EVALUATION
cara.jackson@urbanteachers.org

KIRSTEN MACKLER
DIRECTOR OF PERFORMANCE & EVALUATION
kirsten.mackler@urbanteachers.org

URBAN TEACHERS
1500 Union Avenue, Suite 2200, Baltimore, MD 21211
**Tel** 410-662-6600  **Fax** 410-662-5500
www.urbanteachers.org