Developing Reliable Classroom Observation Systems

Paper presented at the annual meeting of the Association for Education Finance and Policy

Cara Jackson  cara.jackson@urbanteachers.org
URBAN TEACHERS
Abstract

Recent state and federal policies have focused a great deal of attention on teacher evaluation systems as a mechanism to enhance teacher effectiveness and, subsequently, student outcomes. Classroom observation data are especially relevant to leaders of educational agencies given the widespread applicability and availability of these data; however, the reliability and validity of classroom observation data may vary depending upon who conducts the observation. In this study, we examine whether paired observations increase inter-rater reliability and explored how the number of indicators and observations influence the reliability of classroom observation scores. Using data from inter-rater reliability exercises conducted prior to and after the paired observation, we explored whether coaches’ alignment to master ratings improved after the paired observation. In addition, we drew on data from 40 first-year teachers to conduct a generalizability and decision study to determine the impact of the number of indicators and observations on score reliability. The findings guided Urban Teachers’ policy decisions regarding the use of paired observations and the number of observers to use when conducting observations in a high-stakes environment.

I. INTRODUCTION

Recent state and federal policies have focused a great deal of attention on teacher evaluation systems to improve instruction. For example, in 2012, President Obama granted waivers from No Child Left Behind (NCLB) requirements to several states based on their progress toward implementing reform measures that included the development and implementation of rigorous teacher evaluation systems. While teacher evaluation systems are not new in most states and districts, the implementation of more rigorous systems—as well as the increased stakes attached to them—has led to the need for more research regarding the reliability and validity of these measures and effective interventions to ensure the quality of the data.

While evaluation systems have increasingly incorporated multiple measures, such as student survey data and student growth, classroom observations remain a cornerstone of teacher evaluation. Required by 48 of the 50 states (Doherty & Jacobs, 2015), observations can be conducted on any teacher, whereas other measures, such as student surveys and student achievement gains, may not be feasible to implement in certain settings. Compared to measures of teacher performance based on
student achievement gains, observations provide more timely information for principals seeking to
make decisions about which teachers to retain (Goldring et al., 2015).

Perhaps most importantly, some recent studies indicate that classroom observations can
support mid-year corrections in teaching practice, making this measure particularly useful for the
“feedback and support” aspect of teacher evaluation policies. Steinberg and Sartain (2015) examined a
teacher evaluation program designed to increase student learning by improving classroom instruction
through structured principal–teacher dialog; leveraging the staggered nature of the rollout, they found
that at the end of the first year, the schools that implemented the program during the first year
performed better in reading and math than the schools that delayed implementation until the second
year. In Taylor and Tyler’s (2012) study of mid-career math teachers, the authors find evidence that a
quality classroom-observation-based evaluation and performance measures can improve mid-career
teacher performance both during the period of evaluation and in subsequent years.

In Figure 1, we portray the framework for how classroom observation data leads to improved
student outcomes. Observations enable school leaders and instructional coaches to provide targeted
feedback and support to teachers and to selectively retain their more effective teachers. If feedback and
support are effective, and schools are able to strategically retain more of their effective teachers and/or
dismiss ineffective teachers, we would expect to see teaching practice improve on average. This in turn
is expected to lead to improved student outcomes. To succeed in this ambitious goal, the measures
must provide accurate information that differentiates among teachers. If the measures are not valid and
reliable, we risk providing inaccurate feedback, inefficiently targeting support, retaining ineffective
teachers, or dismissing effective teachers. In other words, the potential for observation data to improve
educational outcomes for students depends on the quality of the underlying data, regardless of whether
the system is primarily used for feedback and support or for high-stakes retention decisions.

II. Background

Classroom Observations as a Measure of Teacher Performance

While classroom observations have considerable potential to foster improvements in teaching if
they provide reliable and valid information, Davis, Pool, and 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. Since reliability and validity are functions of the
users of classroom observation tools, as well as of the tool itself (Sartain, Stoelinga, & Brown, 2011), ensuring the quality of classroom observation data requires training and monitoring of use. For example, in the Measures of Effective Teaching (MET) project, researchers ensured observation tools were used consistently by requiring raters to undergo 17 to 25 hours of training and 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. While MET researchers found that scores from various observation instruments were positively associated with student achievement gains under the training and monitoring conditions described above, they also found that achieving high levels of reliability is challenging (Kane & Staiger, 2012).

Studies drawing the MET project shed light on why and how achieving high levels of reliability is challenging. Jerald (2012) notes observers often struggle to transfer their new skills to the field if they have only practiced using videos; one district partner developed “live observation training” as one of the activities provided to support observers. Drawing on data from both the MET and the Understanding Teaching Quality (UTQ) studies, Bell et al. (2014) examined think aloud data to understand how observers and master observers assign scores on observation protocols and learned that observers found it easier to score low-inference dimensions and more difficult to score high-inference aspects of teaching, such as indicators related to interactions among teachers, students, and subject matter. In addition, they found that observers and master observers use scoring strategies (reviewing scoring criteria, using internal or personal criteria, reasoning from memorable videos, and beginning with an assumed score) differentially. Based on their work, researchers recommend providing more time for “live practice” in classrooms (Jerald, 2012) and training differentially, with more time devoted to high-inference dimensions and those with lower levels of accuracy and consistency (Bell et al., 2014).

In addition to training, researchers have used data to explore how to design observation systems in ways that enhance the reliability of classroom observation data. Ho and Kane (2013) examined classroom observation data from the MET project to explore the implications of different approaches to obtaining high levels of accuracy and reliability in classroom observations. Drawing on data from 129 raters (53 school administrators and 76 peer raters) who observed lessons from 67 teachers, they found that even when observers are well trained, a single observation conducted by a single observer is a fairly unreliable estimate of a teacher’s practice (Ho & Kane, 2013). The authors found a 14 percent increase in reliability when two lessons rather than one were observed by the same
rater, and a 32 percent increase in reliability when two lessons were observed by two different raters compared to one lesson with one rater. Hill, Charalambous, and Kraft (2012) contend that these types of studies, which employ generalizability theory as framework for making considering multiple elements of observational systems, can provide empirical evidence regarding the optimal design required to produce desired reliabilities.

_Urban Teachers_

Urban Teachers, which launched as a non-profit organization in September 2009, is a residency-based teacher preparation program designed to provide high-need schools with effective teachers through a combination of rigorous selection, intensive support, and ongoing accountability. In its first year, Urban Teachers opened with 39 residents in Baltimore City Schools and charter management organizations in Washington, DC. In the 2014–15 school year, Urban Teachers welcomed 112 residents for its fifth and largest class yet. In total, 307 residents and teachers serve in 86 schools across Baltimore City and Washington, DC. Urban Teachers’ inaugural class was prepared in general and special education for literacy and math content areas in grades prekindergarten through nine, leading to licensure in both general and special education for all participants who successfully progress through Urban Teachers’ program. Urban Teachers has since launched secondary math and secondary English programs of study.

Participants 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 certification in both the content area and in special education. Along the way, participants undergo ongoing evaluation of teaching practice and growth mindset and professionalism. Urban Teachers’ theory of action for ensuring effective teachers, encompassing selection, support, and evaluation, is visually displayed in Figure 2.

Ongoing evaluation of the performance of teacher candidates, as a basis for feedback and support as well as retention decisions, is a key component of Urban Teachers’ residency-based
preparation program. 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 teacher candidates are eligible for certification, Urban Teachers has gathered multiple sources of data on performance to use as the basis for making that decision.

Urban Teachers’ Teacher Practice Rubric

During classroom observations, observers use the Teacher Practice Rubric (TPR) to assess how well participants are implementing specific teaching practices. The TPR is aligned with the standards set by the Council of Chief State School Officers’ Interstate Teacher Assessment and Support Consortium and encompasses four sets of skills that new teachers must master in order to become effective. All of 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. The teaching practice portion of the SY 2014-15 rubric consists of 19 indicators; see Appendix A for details. Observers rate each indicator on a 0 (no evidence) to 4 (mastery) scale, and the final observation score is the average across all 19 indicators.

Given that ongoing evaluation plays an important role in Urban Teachers’ theory of action, the organization seeks to assess and improve the reliability of performance measures to ensure consistency in feedback and to validate the use of these measures to inform high-stakes decisions regarding whether participants continue in the program. Recognizing the need to ensure uniform application of classroom observation rubrics, Urban Teachers has adopted several of the practices used in the MET project. To this end, Urban Teachers’ Curriculum and Professional Development team holds faculty institutes at the start of each semester include activities to develop inter-rater reliability to train and calibrate coaches on the Teacher Practice Rubric before they begin classroom observations. Prior to the beginning of the school year, instructional coaches complete calibration exercises to assess inter-rater reliability. Following these exercises, Urban Teachers generates individualized reports to share with coaches; these
Even when observers are well trained, a single observation conducted by a single observer is a fairly unreliable estimate of a teacher’s practice (Ho & Kane, 2013). In an effort to identify ways to enhance reliability, we focused on two questions: 1) can paired observations increase inter-rater reliability, and 2) how do the number of indicators and observations influence reliability of classroom observation scores?

III. DATA & METHODS
Paired observations took place throughout fall of 2014. A lead clinical faculty member worked with new instructional coaches and with coaches whose inter-rater reliability was low based on the calibration exercises conducted in August 2014. In total, nine coaches plus the lead clinical faculty member participated in the paired observation pilot. During a paired observation (typically about 45 minutes), the lead clinical faculty member and the teacher’s instructional coach observed a lesson together, taking notes as they observed. Immediately following the observation, the lead clinical faculty member and the coach independently rated various indicators of teaching practice based on their interpretation of the evidence. After completing independent ratings, the pair was asked to compare their ratings and resolve discrepancies through discussion.

To address the first research question, we attended four of the paired observation sessions, documented the paired observation process, and collected data on the discussion of how coaches arrived at specific ratings and how the pairs resolved discrepancies. We also compared inter-rater reliability from the calibration exercises that took place before the paired observations, in August 2014, to after the paired observations, in January 2015. Data sources included field notes from the paired observations and the coaches’ discussions of ratings, as well as the data on inter-rater reliability from calibration exercises from 9 coaches before and after the paired observations.
To address the second research question, in March of 2015, we conducted a generalizability study to assess multiple elements of the observational system, which inter-rater agreement measures cannot do (Hill, Charambolous, & Kraft, 2012). This study provides empirical evidence regarding how the number of indicators rated and the number of lessons observed affected reliability of the score. We constructed a data file that contained ratings on all observations that took place during the 2013–14 school year and generated a subset of observations to analyze. The analytic sample consisted of first-year teachers for whom we had at least three observations in which all 19 indicators were rated. After establishing the proportion of variance attributable to participants and indicators, the variance components were used in the decision study to shed light on how the number of indicators and observations affected reliability of the overall score.

The model implied by the data collection design under the tenets of Generalizability Theory is as follows:

\[ X_{pioc} = \mu + \nu_p + \nu_i + \nu_{pi} + \nu_{o|p} + \nu_{io|p,e} \]

Where \(\mu\) is the grand mean;

\(\nu_p\) is the person effect;

\(\nu_i\) is the indicator effect;

\(\nu_{pi}\) is the person by indicator interaction;

\(\nu_{o|p}\) is the observation by person interaction;

\(\nu_{io|p,e}\) is the person by indicator by observation interaction, confounded with error.
IV. FINDINGS

Research Question 1: Can Paired Observations Increase Inter-Rater Reliability?

Field notes and data collected during the paired observation process provided insights as to how coaches applied the Teacher Practice Rubric and assigned ratings to specific indicators. Following the observation and time spent independently rating the lesson, the lead clinical faculty member and coach discussed the lesson they had just observed. In these discussions, the lead clinical faculty member and coach described physical evidence, such as the presence of anchor charts, as well as statements the teacher made or student behaviors they had observed. As they discussed their ratings on specific indicators, both the lead clinical faculty member and the coaches frequently described what they would have needed to see to give the participant a higher rating.

When initial ratings were in conflict, the lead clinical faculty member consistently cited evidence to support her positions, and referred back to the scoring criteria in the rubric. The coaches cited evidence from the lesson observed, but some also discussed knowledge of the teacher’s practice or behavior outside of the lesson observed. Most coaches referred back to the scoring criteria on the rubric. In some cases, they did not refer back to the rubric initially, but did after seeing the lead clinical faculty member do so. We observed that for some of the indicators, the language of the rubric made it difficult to come to consensus, as it appeared to reflect more than a single dimension of teaching practice.

The process was highly regarded by the coaches participating in the pilot, who viewed the pairings as an effective form of professional development and appreciated the opportunity to have a one-on-one discussion with lead clinical faculty regarding how to interpret the evidence in light of the descriptions in the Teacher Practice Rubric. Though pairing did not result in universal improvement in inter-rater reliability, post-intervention inter-rater reliability averaged about 11 percentage points higher than pre-intervention inter-rater reliability, and gains were greater for coaches with low reliability prior to participating in the paired observations. The six unnormed coaches averaged an increase of 14 percentage points, from 44 percent exact agreement in August 2014 to 58 percent in January 2015.

Figure 3 displays the inter-rater reliability of the nine coaches that participated in the paired observations. The white bars represent the percent of exact agreement based on the average of three inter-rater reliability exercises at the August Institute, while the black bars represent the raters’ exact
agreement with the master rating based on the January 2015 inter-rater reliability exercise. Of the nine coaches who participated in paired observations, seven had a higher proportion of exact agreement with the master rating in the January inter-rater reliability exercise compared to their exact agreement in the August exercises. Coach C remained about the same, and Coach I performed somewhat worse in the January exercise compared to the August exercises.

**Research Question 2: Reliability of Observation Scores under Different Numbers of Indicators and Observations**

We conducted a generalizability study to decompose variability teachers’ observation scores into meaningful components of variance. This was followed by a decision study to determine how reliability is affected by the number of indicators in the Teacher Practice Rubric and the number of observations conducted. The decision study was also to identify scoring protocols that will maximize precision for minimal cost. These analyses are based on a subset of the observations conducted during the 2013-14 school year. Specifically, we used data from 40 teachers with at least three observations that included scores on all 19 indicators.

We first determined the coefficient alpha for each observation. Coefficient alpha is .96 for the first observation, .94 for the second observation, and .97 for the third observation. Thus, the estimated expected correlation between two replications of the measurement procedure, where indicators like these indicators are randomly drawn and administered, is between .94 and .97. We estimate that 94 percent or more of the observed score variance can be accounted for by true score variance.

The correlation of average scores from different observations ranges from .71 to .81, as displayed in Table 1. This is the test-retest reliability, an estimate of the reliability when we consider indicators as fixed and occasions as random. It is an estimate of the expected correlation between two replications across different occasions if indicators are held constant.

In Table 2, we list the sources of variance, the amount of variance, the standard error, and the percent of total variance explained by each source. The $p$ variance is the estimated “true score” variance, or the “good” variance that distinguishes among teachers. The $i$ variance is the variance of individual indicator difficulties; it is neutral variance as long as everyone is rated on the same indicators, and as long as only relative (not absolute) position on the scale matters.
The pi variance is the variance of person-indicator interactions, and error. The o|p variance component refers to the variance of observation-person interactions, and error. These variances reflect the extent to which particular people score higher or lower on particular indicators (or observations) above and beyond what is predicted by person ability and indicator (or observation) difficulty. Both pi and o|p variance are undesirable with respect to relative error. The o|p variance is undesirable because the relative position of the teacher’s score would change if they were observed on a more or less favorable occasion. The pi variance is undesirable because if we sampled different indicators, teacher rankings (relative position) would change. However, because Urban Teachers standardizes indicators across persons (that is, all participants are observed on the same indicators), the pi variance is neutral with respect to relative error in this case.

It is not necessarily the case that greater percentages indicate more important sources of variance, since the variance components are estimated variances of the distributions of single effects. Reported scores are not based on a single indicator, but averages of indicators (and possibly, averages of observations). Thus, the results of the decision study (presented in the main text) describe the importance of each source of error in terms of their impact on reliability.

The findings from decision study indicate that adding observations and rubric categories produces diminishing returns to reliability. As seen in Figure 4, reliability improves most markedly (from .70 to .81 at ten indicators) when increasing from one to two observations; adding a third observation increases reliability to .85, and a fourth increases reliability to .87. Similarly, when we increase the number of indicators from two to four (at two observations), reliability increases from .63 to .73, but reliability only increases from .79 to .81 when we increase the number of indicators from eight to ten.

V. LIMITATIONS

There are a few caveats regarding the results of the pair observation intervention. First, we had a very small sample for our pilot intervention. It remains to be seen whether paired observations would continue to work well among a larger group of instructional coaches or with different lead clinical faculty member guiding the process. Second, because we did not randomly assign a control group of unnormed coaches to a control condition of no treatment, we cannot be certain that those coaches improved due
to the paired observations. It is possible that unnormed coaches would have improved considerably even without this intervention.

The results from the generalizability and decision study are useful to guide decisions regarding the number of observations and number of indicators. However, in this study we explore just two sources of variation. One limitation of this study is that we cannot partition out the variance that is due to the observer, because participants were observed by their instructional coaches and in most cases, the coach remained the same throughout the year. In addition, scores from observation rubrics may vary as a function of the lesson, subject, and grade level.

VI. DISCUSSION

Ensuring the accuracy of observation scores is essential if these data are to be useful for the purposes of helping teachers improve and making high-stakes decisions about teachers. In this study, we pursued two avenues for improving score reliability: a pilot intervention to enhance inter-rater reliability, and a variance decomposition to guide decisions about the number of observations we rely on to inform decisions about which participants continue in the program. Both approaches have yielded valuable insights that have informed our evaluation model.

Assigning new coaches or those who are struggling to perform well on calibration exercises to work with an experienced partner may help them learn to apply effective strategies. The paired observations gave coaches opportunities not only to experience “live practice” as Jerald (2012) recommended, but to do so alongside a lead clinical faculty member who modeled the strategy of referencing scoring criteria in relation to evidence when determining ratings. While observers may use numerous reasoning strategies to arrive at a score, referencing the scoring criteria may enhance rater agreement. Furthermore, the paired observations provided an opportunity for coaches to spend more training time on higher-inference aspects teaching, as recommended by Bell et al. (2014). The paired observations with a lead clinical faculty member were viewed by coaches as an effective form of professional development and yielded promising improvements in inter-rater reliability. In piloting the paired observation program, Urban Teachers strategically selected coaches who either had low exact agreement during the training sessions or were recent hires. Although there is interest in expanding the paired observations to all coaches, Urban Teachers may need to continue to target specific coaches, since this is a resource-intensive approach to enhancing inter-rater reliability. Further research could
explore whether paired observations are more cost-effective in some cases than others. For example, it seems likely that unnormed coaches will benefit more from the paired observations than normed coaches.

The data from the paired observation pilot highlighted specific indicators on the rubric that were difficult for the pairs to rate consistently, even when they referred back to the language of the rubric. We realized the language in the scoring criteria appeared to reflect more than one dimension of teaching practice. Prior to the current school year, we revised the Teacher Practice Rubric in an effort to streamline indicators that were comprised of multiple dimensions of teaching practice, consistent with Bell et al. (2014)’s recommendation to revise scoring criteria to improve clarity and reduce cognitive demand on observers as a means of improving rater agreement. In the current school year, all clinical faculty agree that the Teacher Practice Rubric is a useful tool for evaluating participants, up from 90 percent last year, and the percent that strongly agree that the rubric is a useful tool for evaluating participants increased from 17 percent to 36 percent.

The generalizability and decision studies provided empirical evidence to inform decisions regarding the number of indicators rated and lessons observed needed to achieve desired reliability of observation scores. Because we see a relatively big increase in reliability when we move from one to two observations, Urban Teachers now uses the average from two observations for summative evaluations of our participants. In addition, we use two different observers, in light of Ho and Kane (2013)’s findings that using different raters enhances reliability. Not only does this approach improve reliability of the observation score, but it also provides data that we have used to continually monitor the observation process. Having two sets of observation scores gives us greater insight into whether some observers tend to rate stringently, leniently, or exhibit a high level of discrepancy from other observers. We have used this information to provide differentiated professional development to coaches throughout the year.

We plan to refine our evaluation system as we receive additional feedback from clinical faculty and participants, and are committed to continually monitoring the observation process to ensure data quality. We anticipate revisiting the generalizability and decision studies with the revised rubric, and specifically exploring the proportion of variance attributable to raters. Finally, we are exploring ways to use video during the observation process to ensure that ratings are evidence-based and that coaches are providing useful feedback to participants to support improvements in teaching practice.
References


FIGURE 1. Data Use Framework Applied to Classroom Observation Data
FIGURE 2. Urban Teachers’ Theory of Action
FIGURE 3. Exact Agreement between Coaches and Master Ratings, August 2014 and January 2015
FIGURE 4. Generalizability Coefficient for Absolute Error
Tables

**TABLE 1**
*Correlation between replications*

<table>
<thead>
<tr>
<th></th>
<th>Observation 1</th>
<th>Observation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation 2</td>
<td>0.788</td>
<td></td>
</tr>
<tr>
<td>Observation 3</td>
<td>0.709</td>
<td>0.811</td>
</tr>
</tbody>
</table>

**TABLE 2**
*Variance, Standard Error, and Proportion of Variance Explained*

<table>
<thead>
<tr>
<th>Source</th>
<th>$\hat{\sigma}_p^2$</th>
<th>$\hat{\sigma}_i$</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.442</td>
<td>0.665</td>
<td>39.2%</td>
</tr>
<tr>
<td>i</td>
<td>0.118</td>
<td>0.344</td>
<td>10.5%</td>
</tr>
<tr>
<td>pi</td>
<td>0.119</td>
<td>0.345</td>
<td>10.5%</td>
</tr>
<tr>
<td>o</td>
<td>p</td>
<td>0.132</td>
<td>0.363</td>
</tr>
<tr>
<td>io</td>
<td>p,e</td>
<td>0.318</td>
<td>0.564</td>
</tr>
</tbody>
</table>
## Appendix: Teacher Practice Rubric Strands A-D and Indicators

<table>
<thead>
<tr>
<th><strong>Strand A: Build A Productive &amp; Nurturing Classroom Environment</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1. Physical Environment: Arranges the physical environment to facilitate student engagement in learning and participation through visual, auditory, and movement opportunities.</td>
</tr>
<tr>
<td>A2. Teacher Presence: Has a strong teacher presence in the classroom and uses a strong teacher voice to convey intent and purpose.</td>
</tr>
<tr>
<td>A3. Classroom Management Plan: Implements a classroom management plan that is transparent, consistent, pro-active, fair, and positive.</td>
</tr>
<tr>
<td>A4. Knowing Students: Is aware of student strengths and needs, learning styles, and emotional status and responds quickly and effectively to student behaviors through words and actions.</td>
</tr>
<tr>
<td>A5. Transitions: Instruction is tightly planned.</td>
</tr>
<tr>
<td>A6. Student Ownership: Creates a learner-centered environment that supports students in managing their own learning.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Strand B: Are Diagnosticians</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>B1. Data systems: Knows data – types and relevance. Constructs and maintains data systems to monitor student learning at the small group and whole class level. Has a thorough knowledge of key assessments in literacy and math and implements them routinely.</td>
</tr>
<tr>
<td>B3. Sharing Data: Makes students’ progress transparent to them &amp; their families; uses student conferences to enrich, intervene, or remediate at the individual student level; supports students with “owning” their data and their learning goals.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Strand C: Set Precise Goals &amp; Enact Them</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>C1. Standards: Use knowledge of state and district standards to set benchmarks for student performance.</td>
</tr>
<tr>
<td>C2. Models &amp; Exemplars: Use models/exemplars of work to provide clear expectations for student performance.</td>
</tr>
<tr>
<td>C3. Setting Goals: Plans instruction that engages all students in high-quality, real-world tasks directly related to standards and discipline-specific learning goals.</td>
</tr>
<tr>
<td>C4. Differentiation: Differentiates instruction to meet the needs of all learners.</td>
</tr>
<tr>
<td>C5. Academic Rigor: Continuously push for ongoing development of deep and accurate student understanding.</td>
</tr>
<tr>
<td>C6. Lesson Structure &amp; Pacing: Lesson demonstrates gradual release of responsibility and instruction is appropriately paced, tightly planned, and well-delivered.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Strand D: Foster Academic Conversations</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>D1. Talk Strategies: Deliberately implements strategies to foster academic conversations. Encourages conversations between students.</td>
</tr>
<tr>
<td>D2. Talk Structures: Provides students with clear expectations for academic conversations.</td>
</tr>
<tr>
<td>D3. Orchestrating Academic Talk: Purposefully creates opportunities for students to engage in academic conversations.</td>
</tr>
<tr>
<td>D4. Norms for Talk: Holds all students accountable for participating in respectful academic conversations.</td>
</tr>
</tbody>
</table>