The Essential Elements of Evaluation: Field Testing Observation Rubrics

Monica W. Jordan and John Barker

TEI Paper No. 2

Memphis City Schools is an intensive partnership site for the Bill and Melinda Gates Foundation in a joint project to improve teacher effectiveness. The language and thought of Memphis’ Teacher Effectiveness Initiative (TEI) has been cited widely in the national conversation about teacher effectiveness and improved learning opportunities for students. This paper and the series of papers that will follow describe critical components of TEI implementation in order to share the lessons learned in Memphis with the local, state, national, and international education community.

Each paper is intended to be a briefing report that provides both high-level thoughts and actionable details relative to the work. Follow-up contact with the authors is welcomed and encouraged.

The Context

Research has shown that teachers are the most important school-related factor that influences student achievement. Because what teachers actually do in their classrooms is causal for student results, it is imperative to explore teacher instructional practices when trying to determine what makes teachers effective. Key to this process is ensuring the use of valid and reliable observation tools, observers, and supporting systems. To assist with this investigation, a Teacher Evaluation Working Group was established in the Memphis City Schools with a goal of finding an observation tool with the ability to isolate teaching behaviors that impact student achievement.

The MCS Office of Teacher Effectiveness Measurement (TEM) “road tested” observation rubrics to prepare evidence for discussion at the Teacher Evaluation Working Group’s retreat in early June 2010. A total of thirty-two (32) classrooms were observed during the last two weeks of the 2009-10 school year alternately using one of three rubrics: IMPACT (from Washington, D.C. Public Schools), Teaching as Leadership (from Teach for America), and Charlotte Danielson’s standard observation rubric. The classrooms represented teachers with high, medium, and low value-add scores (based on calculations for our district by Mathematica Policy Research) in each grade span, elementary to high school. Observations were unannounced and lasted approximately 15 to 30 minutes each. In cases where appropriate, teachers, students, and administrators were able to answer interview questions related to instruction, observation, and evaluation after the observation.

Among the rubrics there were several common domains that can be categorized as Planning and Preparation for Learning; Classroom Management; Delivery of Instruction; Monitoring, Assessment, and Follow-up; Family and Community Outreach; and Professional Responsibilities. Each of the domains was

---

scored as accurately as possible given the limitations of time and instructional objectives appropriate for the last two weeks of the school year.

The Results

Prior to the field test, the observation rubrics were analyzed in terms of functionality in a real-time observation, e.g., how many pages, orientation of the page, quantification of each item on the observation, continuum of effectiveness, specificity of language defining instructional practice. A further analysis was performed to investigate the degree to which coverage of the existing domains in the Tennessee Framework could be ascribed to each observation rubric. Each rubric adequately addressed the behaviors within each domain; however, rubrics employ different naming conventions to define the domain and explore the behaviors with varying degrees of depth, resulting in varying levels of sensitivity among the rubrics. Findings from rubric comparisons are presented below.

<table>
<thead>
<tr>
<th>Model</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teach for America, Teaching as Leadership</td>
<td>Descriptive teacher instructional behaviors</td>
<td>Ambiguous language, several pages of standards, difficulty assigning value because rubric lacks concrete examples, required significant inferences to be made in order to score each behavior</td>
</tr>
<tr>
<td>Charlotte Danielson, Framework for Teaching</td>
<td>Descriptive teacher instructional behaviors, focused on the teacher’s actions</td>
<td>Ambiguous language, several pages of standards, difficulty assigning value because rubric lacks concrete examples, required significant inferences in order to score each behavior</td>
</tr>
<tr>
<td>District of Columbia Public Schools, IMPACT</td>
<td>focused on the teacher’s actions and resulting students’ behavior, ease of use in terms of navigation through the rubric, format of the instrument, specificity of language, examples of teacher and or student behavior to guide observation, significant depth with clear, specific, measurable examples</td>
<td>Several pages of standards</td>
</tr>
</tbody>
</table>

The TEM team found the IMPACT rubric by DCPS to have appropriate language, specific descriptors, clearly articulated levels of performance, and ease of use during field testing.

Based on the results from the “road test”, the observation data were then compared to teacher value-add data to identify connections among value-add scores and instructional practices identified by the various rubrics. Preliminary comparison of Mathematica value-add data and observation data collected during the “road test” using the various observation rubrics suggests a pattern of teachers with higher value-add scores having higher ratings in 10 of the 14 standards observed.

<table>
<thead>
<tr>
<th></th>
<th>Teach for America, Teaching as Leadership</th>
<th>Charlotte Danielson, Framework for Teaching</th>
<th>District of Columbia Public Schools, IMPACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Most of the teachers exhibited most of the 33 characteristics</td>
<td>Most of the teachers exhibited most of the 34 characteristics</td>
<td>Most of the teachers exhibited most of the 14 characteristics</td>
</tr>
<tr>
<td>Medium</td>
<td>Many of the teachers exhibited some of the 33 characteristics</td>
<td>Many of the teachers exhibited some of the 34 characteristics</td>
<td>Many of the teachers exhibited some of the 14 characteristics</td>
</tr>
<tr>
<td>Low</td>
<td>Some of the teachers exhibited some of the 33 characteristics</td>
<td>Some of the teachers exhibited some of the 34 characteristics</td>
<td>Some of the teachers exhibited some of the 14 characteristics</td>
</tr>
</tbody>
</table>
Throughout the 2009-10 school year, the district wide Teacher Evaluation Working Group began reviewing observation tools to field test during the 2010-11 school year. The process included guidance from The New Teacher Project (TNTP) on observation rubric selection followed by specific presentations on several highly regarded rubrics, e.g., Charlotte Danielson’s Framework for Teaching, Classroom Assessment Scoring System (CLASS), the Teaching as Leadership (TAL) rubric developed by Teach for America, Tennessee Framework for Evaluation, IMPACT from District of Columbia Public Schools and Teacher Advancement Program (TAP). The committee, including teachers from both tested and non-tested grades and subjects, as well as principals, held a retreat on June 3-4, 2010, where participants reviewed rubrics and provided feedback on each rubric’s ease of use, language specificity, and alignment of domain components to the committee’s perceptions of effective teaching.

Individuals on the committee were asked to prepare final presentations on each rubric to remind the group of learning that had taken place earlier in the year on previously presented rubrics. Presentations were made followed by discussions of the advantages and disadvantages of each rubric in regard to ease of use, language specificity, and alignment of domain components to the group’s perceptions of effective teaching. Of note, the committee appreciated the student focus that the IMPACT rubric emphasized for the observer. From there, the committee selected three rubrics to field test in 2010-11: 1) the revised MCS observation component, 2) IMPACT, and 3) the TAP rubric.

The Implications

Vital to a robust evaluation system is an observation model that enables the observer to isolate specific teacher moves and behaviors which lead directly and undeniably to student achievement. A valid observation instrument and reliable observer operating within a supportive system are critical components which require a thoughtful and well-organized implementation. Several broad strategic decision points arise:

How will teachers be observed? Possible scenario: Teacher observations will occur both in person and through video. Rubrics will be field tested in 2010-2011 to determine appropriate rubric(s) for teachers in varying grades and subjects.

Who will conduct observations? Possible scenario: Teachers (TVAAS and Non-TVAAS) will be observed by principals and peer observers. Regional teams (regional superintendents, instructional coaches, etc.), Content-Specific Coordinators (Music, Art, World Language, Orff, ESL, SPED, Career and Technology, Physical Education and Lifetime Wellness, ROTC, etc.) and a newly hired team of Teacher Effectiveness Measurement (TEM) observers are included in the category of peer observers. These observations occur in person or through review of video by observers trained by the district through a state-approved training program.

When will observations occur? Possible scenario: Teachers will be observed four times annually, occurring once per nine-week grading period. Announced observations are not required. In order to conduct four observations (1 per quarter) to 70,000 teachers statewide, a total of 260 observations must occur per hour (based on 6 hours in a school day for 45 instructional days per quarter). In Memphis City Schools, in order to conduct 4 observations per quarter for 7,000 teachers, approximately 26 observations must occur per hour.
Which technology is needed to collect and to convey observation data? Possible scenario: Electronic versions of each rubric will be available to use with a hand-held device, allowing for instant population of observation data to a data warehouse. Through this data repository, results will feed back to teachers for reflection, to principals for follow-up, to the TEM system for annual evaluations, and to teacher support staff to ensure teacher support and growth.

What actions are needed to ensure observations occur? Possible scenario: Districts must self-examine to identify current practices (classroom walkthroughs by district staff, drop-ins by principals, etc.) that facilitate or threaten observations. MCS, like other districts, must figure out how to adequately address the capacity issue that presents itself when attempting to evaluate 7,000 teachers annually.

Please direct requests for additional information on this topic to Dr. John R. Barker at barkerjohn@mcsk12.net