



A Comparison of Mathematics Classroom Observation Protocols

Melissa Boston
Duquesne University

Jonathan Bostic
Bowling Green State University

Kristin Lesseig
Washington State University—Vancouver

Milan Sherman
Drake University

In this article, we provide information to assist mathematics teacher educators in selecting classroom observation tools. We review three classroom observation tools: (1) the Reform-Oriented Teaching Observation Protocol (RTOP); (2) the Instructional Quality Assessment (IQA) in Mathematics; and (3) the Mathematical Quality of Instruction (MQI). We begin by describing each tool and providing examples of research studies or program evaluations using each tool. We then look across tools to identify each tool's specific focus, and we discuss how the features of each tool (and the protocol for its use) might serve as affordances or constraints in relation to the goals, purposes, and resources of a specific investigation. We close the article with suggestions for how each tool might be used by mathematics teacher educators to support teachers' learning and instructional change.

Key words: Classroom observation tools; Instructional quality; Reform-oriented teaching

In the September *Mathematics Teacher Educator* editorial, Smith (2014) describes how "tools" originally developed for research can be utilized by mathematics teacher educators to support teachers' learning and instructional

change. Smith provides examples of how research frameworks for analyzing instructional tasks and teachers' questions can serve as scaffolds for teachers' learning (about cognitive demands or question types), instructional practice (in selecting tasks or asking questions), and reflection (on the nature of tasks or questions used during a lesson). In these examples, and more generally, tools highlight particular aspects of practice that research has identified as critical for enhancing students' learning of mathematics. Through the lens of a specific tool, teachers may be able to see aspects of instruction that previously blended into the myriad classroom activities occurring throughout a lesson. Once aspects of instruction are made visible, tools can provide a concrete structure for the development of new practices by specifying criteria and identifying standards for the implementation of the intended practice. Finally, tools can foster formative assessment and self-evaluation by focusing teachers' reflections on emerging or existing practices to identify strengths and/or motivate change.

In mathematics teacher educators' work as researchers, tools can serve to focus our analysis on key features of an intervention or treatment (e.g., professional learning activity, professional development initiative, or teacher education course or program). Tools can communicate standards for components of practice (e.g., mathematical tasks or teacher's questions) that can be shared across institutions, programs, research groups, and professional development settings. Tools allow us to gather data specifically related to the question under investigation (Smith, 2014), generating evidence that directly indicates the effectiveness or impact of the intervention and enables us to make valid claims about the intervention. In this way, tools can support evidenced-based practice in mathematics teacher education, particularly when the same tools are employed across multiple projects, sites, and investigators.

In this article, we take a closer look at a specific set of research tools that also hold promise for supporting teachers' learning and instructional change—classroom observation instruments. When the purpose of a pro-

This is an invited manuscript. In accordance with journal procedures, the manuscript was reviewed by the editor and two members of the Editorial Panel. It did not undergo a double-blind review process.

This manuscript is based on a presentation at the Association of Mathematics Teacher Educators (AMTE) 2014 Annual Meeting. The analysis of classroom observation tools described in this manuscript began within a working group in the Service, Teaching, and Research Program (STaR) in Mathematics Education in which the second, third, and fourth authors participated. Each of these authors contributed equally to this manuscript, and they are listed alphabetically.



professional development project, intervention, or teacher education program is to change or impact some aspect of teachers' instructional practices, classroom observations are an essential component in identifying whether the intended practice is being implemented by teachers. The tools used to observe and assess classroom instruction must be able to identify the instructional practice(s) under investigation, using valid and reliable measures. Data and results generated by the classroom observation instrument should be useful in evaluating the effectiveness of the project, intervention, or program *and* for providing feedback to inform improvements in teachers' practice.

In this article, we provide information to assist mathematics teacher educators in selecting a classroom observation tool. We review three classroom observation tools and the protocols for their use:¹ (1) the Reform-Oriented Teaching Observation Protocol (RTOP); (2) the Instructional Quality Assessment (IQA) in Mathematics; and (3) the Mathematical Quality of Instruction (MQI). We have purposefully selected three tools with different foci and methods of use in order to raise issues regarding the selection of classroom observation tools more broadly. We begin by describing each tool and providing examples of research studies or program evaluations using each tool. We then look across tools to identify each tool's specific focus, and we discuss how the features of each tool (and the protocol for its use) might serve as affordances or constraints in relation to the goals, purposes, and resources of a specific investigation. We close the article with suggestions for how each tool might be used by mathematics teacher educators to support teachers' learning and instructional change.

The Reformed Teaching Observation Protocol

Purpose and Theoretical Foundation

The Reformed Teaching Observation Protocol (RTOP; Sawada et al., 2002), created originally to gather data for the Arizona Collaborative for Excellence in the Preparation of Teachers (ACEPT), is a classroom observation protocol designed to measure the degree to which mathematics and science teaching are reform-oriented. Sawada and colleagues use the term *reform-oriented* to synthesize three instructional aspects: standards-based teaching,² an inquiry orientation in lesson design and implementation, and student-centered teaching practices. This protocol,

grounded in a constructivist view of teaching (von Glasersfeld, 1989), builds upon reform efforts and advancements in K–12 standards for mathematics and science instruction (American Association for the Advancement of Science [AAAS], 1989; National Council of Teachers of Mathematics [NCTM], 2000; National Research Council [NRC], 1996).

The primary goal for the RTOP was to support reform efforts in professional development and teacher education. This goal demanded an instrument that would not only evaluate mathematics and science teaching but also help improve instruction (e.g., to provide a formative assessment of classroom instruction). In other words, evidence from an evaluation could communicate what went well during instruction, what could be improved, and how this improvement might occur.

The RTOP Instrument

The RTOP is a 25-item Likert-scale questionnaire (sample rubric provided in [Appendix A](#); link to entire RTOP provided in [Appendix D](#)) examining three factors within the learning environment (subscales in parentheses): *Lesson Design and Implementation*, *Content* (propositional and procedural knowledge), and *Classroom Culture* (communicative interactions and student/teacher relationships). A brief description of the factors and subscales is offered here; Figure 1 provides example items to illustrate each subscale.

Lesson design and implementation, the only factor without subscales, identifies ways a teacher designs and sequences a lesson to support meaningful learning. The second factor, *Content*, includes two subscales: (1) *propositional knowledge* assesses whether instruction focuses on understanding core ideas meant to build conceptual understanding, and (2) *procedural knowledge* assesses how students solve problems and engage in problem-solving behaviors. Items on the propositional and procedural knowledge subscales are broad in nature and are not content specific. The third factor, *Classroom culture*, also has two subscales: (1) *communicative interactions* identify discourse moves that occurred during instruction, and (2) *student/teacher relationships* capture teacher moves that facilitate a caring and nurturing environment.

The RTOP uses five-point Likert scales for each item. Raters are provided space for field notes, which supple-

- 1 We use the terms *classroom observation instrument* and *classroom observation tool* synonymously to refer to the set of rubrics used in classroom observations. We use the term *protocol* to encompass the rubrics and the requirements for their use (e.g., rater training, how they are used during an observation, etc.).
- 2 Given the cross-disciplinary nature of the RTOP, “standards-based teaching” is intended to highlight process standards for doing mathematics and engaging in science rather than content-specific standards.

| RTOP factors and subscales | RTOP sample items |
|--|---|
| Lesson Design and Implementation (5 items) | "In this lesson, student exploration proceeded formal presentation." |
| Content | |
| Propositional knowledge (5 items) | "The lesson promoted strongly coherent conceptual understanding." |
| Procedural knowledge (5 items) | "Students made predictions, estimates, and/or hypotheses and devised means for testing them." |
| Classroom Culture | |
| Communicative interactions (5 items) | "The teacher's questions triggered divergent modes of thinking." |
| Student/Teacher relationships (5 items) | "The metaphor 'teacher as listener' was very characteristic of this classroom." |

Figure 1. Examples of RTOP items for each subscale.

ment the quantitative scores and provide evidence to be used for shared reflection with the observed teachers. On the 0–4 scale, a score of 0 indicates that the item was “never observed” during the lesson, a score of 2 indicates that the item was observed at least twice, and a score of 4 indicates that the item was “very descriptive” of the lesson. A total of 10 or greater on any subscale suggests evidence of reform orientation for the construct assessed by that subscale. The RTOP’s internal consistency is exceptionally high, $\alpha = 0.97$, indicating that individual teachers tend to score similarly between subscales. The subscales also have sufficient internal consistency (e.g., teachers tend to score similarly on the five items within a subscale), ranging from $\alpha = 0.80$ to as high as $\alpha = 0.93$, surpassing the threshold for reliable use in research settings (Gall, Gall, & Borg, 2007). Subscales may be used in place of the entire instrument if desired.

Administration of the RTOP

Researchers can use the RTOP to assess videotaped or live mathematics or science instruction in grades K–12, community college, or university settings. More observations of the same teacher increase the statistical validity of the instrument at the teacher level; however, one observation with two or more raters is sufficient to draw conclusions about an instructional episode. Prior to using the RTOP, raters are expected to complete a free online training (link provided in Appendix D), which takes approximately 1–2 days and should be completed by teams of at least two raters.³ The online modules include

(a) understanding the design of the RTOP and (b) reaching adequate interrater reliability with established raters. During the training, trainees evaluate recorded instruction and compare their scores with benchmark scores from RTOP developers. To be considered certified to use the RTOP in formal research, trained raters should have overall scores within ± 5 points of the developers’ scores, with each item score varying 1 point or less (Sawada et al., 2002). In research, two or more trained raters are expected to observe each lesson and then draw agreed-upon conclusions about the observed lesson. Note that the online training videos feature college-level education courses, and additional practice coding videos from K–12 mathematics classrooms may be necessary before using the RTOP.

Interpretation of RTOP Results

The RTOP generates an overall total score for each observed lesson, created by summing scores on all items within subscales. Scores range from 0 to 100, with higher scores indicating greater reform orientation. In a validation study including 141 Grade 6–16 mathematics and science teachers, the average RTOP score was 51.3 with a standard deviation of 20.1 (Sawada et al., 2002). An overall score of 50 (e.g., 25 items \times 2 points per item) is the minimum threshold for considering instruction to have elements of reformed teaching (Sawada et al., 2002). For example, a score of 60 would suggest that the observed lesson was reform-oriented. Deeper inspection of the findings (e.g., by looking at scores for individual factors,

³ In-person training is no longer available.

subscales, or items) would be necessary to determine whether there was a general trend of reform-orientation across all factors or noticeable differences between factors (e.g., a total score of 10 or greater on any subscale suggests evidence of reform-oriented teaching for those specific constructs).

Quantitatively, RTOP scores can be used to assess change over time (e.g., by comparing teachers' pre/post scores), to assess the impact of an intervention or program (e.g., comparing control and treatment groups), or to provide a general measure of reform-oriented teaching in a school or district (e.g., percent of teachers above the 50-point threshold). Qualitatively, RTOP results are intended to foster discussions with individuals or groups of teachers about what raters noted during observed lessons and ways that mathematics or science teaching might be revised to promote reform-oriented teaching practices. In this way, the RTOP can be used in a school or district as a vehicle to spur building-wide or district-wide reform initiatives or in university settings to promote mathematics and science preservice teachers' ideas and practices regarding reform-oriented teaching.

The RTOP as a Research Tool

The RTOP has been used in several studies of mathematics and science teaching (Adamson et al., 2003; Dunleavy, Dede, & Mitchell, 2009; Jong, Pedulla, Reagan, Salomon-Fernandez, & Cochran-Smith, 2010; Roehrig & Kruse, 2005), and it has been adapted to meet specific needs in other studies (e.g., Ciancolo, Flory, & Atwell, 2006; Morrell, Wainwright, & Flick, 2004; Wainwright, Morrell, Flick, & Schepige, 2004). We present two examples here of how the RTOP was used to identify reform-oriented instructional practices.

Examining secondary instruction. Adamson and his ACEPT team (2003), the developers of the RTOP, conducted research to identify the impact of university coursework led by ACEPT instructors on secondary mathematics and science teachers' use of reform-oriented instructional practices. Two research questions guided this study: (1) Was there a difference in RTOP scores between teachers previously enrolled in ACEPT courses and those who did not experience such ACEPT courses? and (2) In what ways did students' content knowledge differ when comparing Grade 6–12 students taught by graduates from the ACEPT program and their peers taught by a non-ACEPT graduate?

ANOVA analyses identified a significant difference in mean RTOP scores between ACEPT graduates and comparison teachers, $F(2,26) = 3.44, p < .05$. Science students

taught by ACEPT graduates had greater content knowledge than their peers taught by non-ACEPT graduates, $F(2,13) = 6.23, p = .01$, as measured by content-specific instruments. Thus, university instruction promoting reform-oriented teaching was associated with better teacher and student outcomes compared to a similar peer group.

Examining elementary instruction. Jong and colleagues (Jong, Pedulla, Reagan, Salomon-Fernandez, & Cochran-Smith, 2010) used the RTOP to examine classroom practices of preservice elementary mathematics teachers and to determine whether student teachers were enacting reform-oriented instruction after experiencing university coursework aligned with the vision of teaching shared by NCTM (2000) and the RTOP authors. Furthermore, the districts where student teachers were placed utilized reform-oriented curricula. The study investigated these questions: (1) To what extent did a sample of elementary preservice teachers implement reform-oriented teaching? and (2) What is the relationship between these teachers' levels of reformed instruction and their students' mathematical understanding (as measured by district-level mathematics tests)?

Results indicated that on average, the elementary student teachers engaged in reform-oriented mathematics teaching. Mean scores for the five subscales ranged from 1.94 to 2.42, which is near or above the minimum threshold (≥ 2) for instruction showing characteristics of reformed teaching. The overall RTOP average was above 50, indicating that in general, student teachers' instructional practices had characteristics of reform-oriented instruction. Subsequent analysis showed that three of the five RTOP subscales (propositional knowledge, procedural knowledge, and student/teacher relationships) were moderately correlated with students' mathematics understanding as measured by the district-level tests (i.e., $r \geq 0.51$). Thus, the student teachers appeared to be enacting the reform-oriented instructional practices advocated by their teacher preparation program, and these practices were associated with increased student learning.

Summary of RTOP

The RTOP has been used in several studies of mathematics and science teaching for 15 years (see Appendix D) and continues to be cited and utilized as a basis for new protocols (e.g., Morrell, Wainwright, & Flick, 2004). Studies using the RTOP share a vision that reform-oriented instruction is likely to lead to productive student outcomes. The RTOP plays an important role by instantiating this shared vision, which then allows observers and teachers to have a common ground for instructional

conversations within and between the content areas of mathematics and science. The RTOP generates quantitative data and also lends itself to offering meaningful feedback that can be conveyed to a teacher as a way to revise his or her mathematics or science instruction to align with the vision of reform-oriented instruction.

The Instructional Quality Assessment Mathematics Toolkit

Purpose and Theoretical Foundation

The Instructional Quality Assessment (IQA) Mathematics Toolkit (Matsumura, Garnier, Slater, & Boston 2008; Boston & Wolf, 2006) is a classroom observation protocol designed to measure the quality of mathematics instruction at scale using a combination of lesson observations, assignment collections, and student work. The IQA is based on two main constructs: *Academic Rigor* and *Accountable Talk*. The primary theoretical framework for Academic Rigor is Stein and colleagues' Mathematical Tasks Framework (Stein, Grover, & Henningsen, 1996), which considers the *cognitive demand* (i.e., type and level of thinking) that a mathematical task can potentially elicit from students, and how cognitive demands change throughout a lesson. Accountable Talk (Resnick & Hall, 1998) consists of the mathematical quality of classroom discourse with respect to accountability to the learning community and to the discipline of mathematics. Thus, the IQA assesses the quality of instruction based on the mathematical work that students *do* and *discuss* in the classroom, based on the cognitive demands and accountable talk moves observed during the lesson.

The IQA Instrument

Figure 2 provides an overview of the IQA constructs, rubrics, and indicators⁴ (sample rubric provided in [Appendix B](#); link to entire instrument provided in Appendix D). The construct of Academic Rigor contains three rubrics. First, *Potential of the Task* identifies the highest level of thinking and explanation that the written task has the potential to elicit from students. Second, *Task Implementation* measures the highest level of thinking in which the majority of students actually engaged during the observed lesson. Third, *Rigor of the Discussion* assesses the level of students' mathematical thinking and reasoning evident during a whole-class discussion following students' work on the task.

The *Rigor of the Discussion* rubric is a holistic assessment of students' mathematical representations, explanations, and strategies provided during whole-class discussion, whereas the Accountable Talk rubrics measure specific elements of classroom discussion at a finer grain size. *Accountable Talk* contains five rubrics: Participation, Teacher's Linking, Students' Linking, Teacher's Press, and Students' Response. *Participation* captures the proportion of students participating in the discussion. *Teacher's Linking* and *Students' Linking* capture accountability to the learning community, measured by the degree to which the teacher or students make connections to and build upon others' contributions to the discussion. *Teacher's Press* assesses the extent to which the teacher requires students to explain and justify their thinking, while *Students' Response* assesses the extent to which students provide such explanations or justifications. These rubrics

| IQA construct | IQA rubric | IQA indicator |
|------------------|-------------------------|---|
| Academic Rigor | Potential of the Task | Instructional Tasks |
| | Task Implementation | Task Implementation |
| | Rigor of the Discussion | Explanations of Mathematical Thinking and Reasoning |
| Accountable Talk | Participation | |
| | Teacher's Linking | |
| | Students' Linking | |
| | Teacher's Press | |
| | Students' Response | |

Figure 2. An overview of the IQA constructs and rubrics.

4 Rubrics for *Rigor of Teacher's Questions* and *Mathematical Residue* were added to the IQA in the fall of 2014 and are not discussed herein.



capture accountability to knowledge and rigorous thinking in the discipline of mathematics.

Each rubric is scored from 0–4 (with 0 indicating the construct is absent). For the Academic Rigor rubrics, low levels of cognitive demand (e.g., memorization or recall of facts or formulas, or the use of previously learned mathematical procedures without connections to concepts or meaning [Stein, Grover, & Henningsen, 1996; Stein, Smith, Henningsen, & Silver, 2009]) correspond to scores of 1 or 2, respectively. High-level cognitive demands (e.g., developing mathematical meaning for or with given procedures and/or open-ended problem solving [Stein, Grover, & Henningsen, 1996; Stein, Smith, Henningsen, & Silver, 2009]) that do not explicitly prompt students to explain their reasoning are assigned a score of 3, while those that do contain such prompts score 4. For the Accountable Talk rubrics, infrequent or formulaic talk moves (or student responses) score 1 or 2, and the presence and consistency of high-level talk moves score 3 or 4, respectively. As few as two observations per teacher may be sufficient to provide a reliable indicator of instructional quality when teachers comply with the data collection requirements (i.e., students engage in mathematical work followed by a whole-class discussion). When teachers were observed on consecutive days by the same (trained) observer, the IQA was found to have an acceptable internal consistency reliability ($\phi = .86$) with only two observations (Matsamura, Garnier, Slater, & Boston, 2008). The dependability coefficient increased to .90 with one additional observation and to .94 for five observations.

Administration of the IQA

The IQA rubrics can be used in observations of K–12 mathematics classrooms. Researchers, professional development providers, and/or teacher educators can select individual IQA constructs or rubrics as aligned with the goals of the project, intervention, or program. A two-day face-to-face training is required for researchers to use the IQA; online training is not currently available. IQA developers intentionally designed the IQA rubrics to be used reliably during live classroom observations, but the IQA can also be used with videotaped lessons.⁵ During the observation, raters take detailed field notes that are used to complete rubrics immediately following the observation. Once pairs of raters have achieved at least 80% exact-point agreement in the field (or using videotaped lessons from a project's dataset), lessons may be observed/scored by an individual rater.

Interpretation of IQA Results

When using the IQA, raters give the observed lesson one score on each IQA rubric. Because the IQA rubrics yield ordinal data, results from across an entire program, school, or project are reported as number (and percentages) of lessons at each score level for each rubric. Comparisons (e.g., between teachers' pre- and post-workshop data, data from control vs. project teachers, or different rubrics) should be conducted using nonparametric tests or tests of frequency or proportion. Descriptive data for each rubric, such as means and medians, are often reported to support interpretations of the results. For example, given the 4-point scale across all rubrics, a mean (or median) score above 2.5 is interpreted as an indicator of higher/consistent use of cognitively challenging tasks and/or accountable talk moves. Further, it is possible to make comparisons and identify relationships across different rubrics (e.g., comparing *Potential of the Task* to *Task Implementation* to determine whether high-level task demands were maintained) because the rubrics' score levels are similarly structured. The IQA score levels are also very descriptive, indicating specific characteristics or frequencies of instructional practice necessary for each score level. The detailed descriptors for each score level and the consistency in score levels across rubrics facilitate qualitative interpretations of the IQA results.

The IQA as a Research Tool

The IQA has been used to assess mathematics teachers' instructional practices in large-scale studies at the school or district level (e.g., Jackson, Garrison, Wilson, Gibbons, & Shahan, 2013; Quint, Akey, Rappaport, & Willner, 2007; Wilhelm, 2014) and professional development research (e.g., Boston & Smith, 2009, 2011; Sztajn, Wilson, Edgington, & Confrey, 2011). In this section, we describe two studies that used the IQA to identify reform-oriented instructional practices in secondary and middle level mathematics classrooms following teachers' participation in professional development or curricular reform efforts.

Measuring the effect of secondary mathematics teacher professional development. Boston and Smith (2009) used the IQA Mathematics Toolkit to assess the effectiveness of professional development designed to help teachers select and implement cognitively demanding tasks. Participants were 18 secondary mathematics teachers in the "Enhancing Secondary Mathematics Teacher Preparation" (ESP) project. Ten secondary mathematics teachers

5 A unique feature of the IQA is that it can be used with collections of student work *in lieu of* classroom observations. As the purpose of this paper is to discuss a sample of *classroom observation* protocols, the rubrics for student work collections are not discussed. See Matsumura, Garnier, Slater, & Boston (2008) for technical quality and Boston (2012) for examples of use.

who did not participate in ESP served as a control group. Instructional tasks, student work, and lesson observations were collected in fall, winter, and spring from ESP teachers and spring only from control group teachers. Only the IQA *Task Potential* and *Implementation* rubrics were used in this study because Accountable Talk was not a central feature of the ESP professional development.

Mann-Whitney tests indicated that ESP teachers' mean *Potential of the Task* score increased significantly from 2.54 to 3.01 from fall to spring ($z = 2.34, p < .01$ [one-tailed]). Chi-squared tests also indicated that the number of high-level tasks ($\chi^2(2) = 16.18; p < .01$) and implementations ($\chi^2(2) = 16.11; p < .001$) in ESP teachers' data collections increased significantly over time. While no significant difference existed between the control group and ESP teachers' fall lesson observations, ESP teachers' scores for spring lesson observations were significantly higher than scores of the control group for *Potential of the Task* ($z = 2.15, p = .02$ [one-tailed]) and *Implementation* ($z = 1.87; p = .03$ [one-tailed]). In addition, the results were independent of the type of curriculum teachers used, providing further evidence of the effectiveness of the professional development in supporting teachers to select and implement cognitively challenging instructional tasks.

Examining the Effect of Standards-based Curriculum and Professional Development. Boston (2012) used the IQA Academic Rigor and Accountable Talk rubrics to examine the effect of a district-wide curriculum adoption and accompanying professional development with 13 middle school mathematics teachers. Although lesson observations and assignments with accompanying student work served as data for the study, we focus here on the analysis of the lesson observations.

Results showed that for lesson observations following teachers' participation in the professional development initiative, the majority of tasks that teachers selected (58%) and implemented (65%) were low level (1 or 2 for *Potential of the Task* and *Implementation*). Only 27% of the whole-class discussions exhibited evidence of high-level thinking and reasoning (3 or 4 for *Rigor of Discussion*), while 54% were considered low level, and 19% of observed lessons lacked any discussion. Results were comparable for each Accountable Talk rubric, with minimal class discussions scoring 3 or 4 for *Teacher's Linking* and *Students' Linking*, or *Teacher's Press* and *Students' Response*.

These results demonstrated that teachers did not seem to be utilizing the high-level tasks provided by their *Standards-based* curriculum on a consistent basis for

classroom instruction. Additionally, teachers were not consistently enacting whole-class discussions, and when they did, the discussion contributed little to students' opportunities to learn. Boston (2012) explained how these, and other, more fine-grained results based on the IQA, can be used to provide specific feedback to district leaders regarding the efficacy of the new curriculum adoption and professional development initiative. Furthermore, the results suggest pathways for improvement in teachers' instructional practice that can be addressed by ongoing professional development tailored to the needs of the district.

Summary of the IQA

The IQA is a holistic assessment of mathematics instruction with a specific focus on the opportunities for students to engage in cognitively challenging mathematical work and thinking and to explain and express their reasoning in whole class discussions. Hence, the IQA rubrics are "best-suited for assessing reform-oriented instructional practices, for use in implementation studies of curriculum or professional development or to identify changes in the nature of school- or district-wide instructional practice over time" (Boston, 2012, pp. 95–96). The small number of observations needed to obtain a stable indicator of classroom practice allows for it to be used at scale and in settings in which videotape is not practical or possible. By identifying and measuring aspects of mathematics instruction that correspond to student achievement (Boaler & Staples, 2008; Stein & Lane, 1996; Stigler & Hiebert, 2004), the IQA provides feedback that can assist mathematics teacher educators in designing professional development that addresses specific elements of instruction (e.g., tasks, implementation, and/or discussion). Thus, the IQA can serve as both a lesson observation protocol and a tool for professional development.

The Mathematical Quality of Instruction

Purpose and Theoretical Foundation

The Mathematical Quality of Instruction (MQI) is a multidimensional assessment of the rigor and richness of the mathematics present during classroom instruction. The MQI must be used on videotaped instructional episodes (rather than live classroom observations). The instrument, designed by the Learning Mathematics for Teaching Project (<http://www.sitemaker.umich.edu/lmt/home>), was originally developed alongside efforts to conceptualize and validate measures of mathematical knowledge for teaching (MKT) (Ball, Thames & Phelps, 2008). MKT refers to the mathematical knowledge that is specifically entailed in the work of teaching. In accord with perspec-



tives advanced in the MKT literature, the MQI is designed to attend to the mathematics-specific components of the lesson and does not preference or measure a particular pedagogical approach. In other words, the authors claim that it is possible and beneficial to attend to the mathematical quality of an instructional episode regardless of the instructional methods. This orientation is reflected in the instrument's attention to *what* rather than *how* mathematical work is evidenced during the lesson.

A second key construct informing the design of the MQI is the instructional triangle (Cohen, Raudenbush, & Ball, 2003) in which instruction is conceptualized as interactions among teachers, students, and content. Given this more encompassing view of instruction, the instrument measures mathematical quality based on what a teacher says and does, what the students say and do, and what the curricula afford.

The original coding scheme was developed and refined through a synthesis of literature on mathematics classroom instruction and analysis of over 250 recorded lessons from 2nd- to 6th-grade classrooms. The codes are intended to capture elements of lessons that compromise the mathematical integrity of a lesson (e.g., the presence of errors or imprecise language), as well as aspects of instruction that support student learning (e.g., the use of multiple representations and explanations that focus on why something works). The instrument was revised in February 2014 to refine codes based on feedback from validation studies and to more explicitly align with mathematical practices outlined in the *Common Core State Standards for Mathematics* (CCSSM) (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010).

The MQI Instrument

The goal of the MQI is to measure the nature of the mathematical content available to students during instruction. To this end, the current instrument is organized around five dimensions of instruction: *Classroom Work is Connected to Mathematics*, *Richness of the Mathematics*, *Working with Students and Mathematics*, *Errors and Imprecision*, and *Common Core Aligned Student Practices* (see Figure 3; sample rubrics provided in [Appendix C](#)). Researchers can decide to use some or all of the MQI dimensions.

The six subscales within the *Richness of Mathematics* dimension capture the extent to which teachers and/or students (1) explicitly link and connect representations of mathematical ideas or procedures; (2) provide mathematical explanations that focus on why rather than how; (3) attend to the meaning of number relationships and

operations; (4) discuss multiple procedures or solution methods; (5) develop mathematical generalizations based on examining instances or examples; and (6) fluently use mathematical language.

The dimension of *Working with Students and Mathematics* captures whether teachers can understand and respond to mathematical ideas students present and appropriately remediate student errors. The *Errors and Imprecision* dimension is more strictly focused on the teacher's use of correct, clear, and precise mathematical language and notation. The final dimension, *Common Core Aligned Student Practices*, incorporates the elements from what was called "Student Participation in Meaning-Making and Reasoning" in previous versions, along with additional subscales to denote the extent to which students work on contextual tasks and communicate about the mathematics. As a whole, subscales in this dimension measure student engagement in sense-making as indicated by the quality of student explanations; evidence of students' questioning, conjecturing, and generalizing mathematical ideas; and the cognitive demand of the task as it is enacted.

Videotaped lessons are chunked into equal intervals of 5 to 7.5 minutes, with each segment coded along the five dimensions. The first dimension, *Classroom Work is Connected to Mathematics*, is simply coded yes or no depending on whether at least 50% of class time (at least 3.75 minutes in a 7.5-minute segment) is connected to mathematics, rather than management or other activities. For the remaining four dimensions, raters take notes on each video segment and use these notes to score a number of subscales and the dimension overall after viewing the entire video (see Figure 3). Subscales and overall dimensions are scored using 4-point rubrics ranging from not present (0) to high (3). To illustrate, the rubric for the subscale *Linking Between Representations* and the dimension *Overall Richness of the Mathematics* are included in Appendix C. Researchers can opt to follow the "MQI Lite" protocol, where they provide an overall score for each dimension but do not score individual subscales.

Administration of the MQI

The MQI is designed for coding video of K–9 classroom instruction in mathematics. Video-based training is available free and online (link provided in Appendix D). Training requires approximately 16 hours that can be parsed into 1- to 2-hour increments. The training modules consist of detailed descriptions of codes and scoring guidelines, along with videos exemplifying different score points and practice tests. After working through these modules, individuals can become MQI certified by achieving an established percentage of agreement with master coding

| MQI dimension | MQI subscales |
|---|--|
| Classroom Work is Connected to Mathematics | 1. Overall Classroom Work is Connected to Mathematics |
| Richness of the Mathematics | 2. Linking Between Representations 3. Explanations 4. Mathematical Sense-Making 5. Multiple Procedures or Solution Methods 6. Patterns and Generalizations 7. Mathematical Language 8. Overall Richness of the Mathematics |
| Working with Students and Mathematics | 9. Remediation of Student Errors and Difficulties 10. Teacher Uses Student Mathematical Contributions 11. Overall Working with Students and Mathematics |
| Errors and Imprecision | 12. Mathematical Content Errors 13. Imprecision in Language or Notation 14. Lack of Clarity in Presentation of Mathematical Content 15. Overall Errors and Imprecision |
| Common Core Aligned Student Practices | 16. Students Provide Explanations 17. Student Mathematical Questioning and Reasoning (SMQR) 18. Students Communicate about the Mathematics of the Segment 19. Task Cognitive Demand 20. Students Work with Contextualized Problems 21. Overall Common Core Aligned Student Practices |

Figure 3. Dimensions and subscales in the MQI.

on a selection of videos. The entire MQI protocol is made available once an individual has completed training and is consider certified.

Interpretation of MQI Results

In order to provide generalizable, reliable indicators of mathematical quality at the teacher level, the developers recommend that at least 3 lessons be scored independently by two coders. As described by Hill, Charalambous, and Kraft (2012), internal consistency reliabilities for overall dimensions substantially increased when three observations (per teacher) were scored by two certified coders compared to just one certified coder. When four observations were scored by two certified coders, internal consistency reliabilities continued to increase, but only slightly; for example, *Richness of the Mathematics*

increased from 0.77 to 0.80, and *Errors and Imprecision* increased from 0.71 to 0.75.

A composite lesson score can be obtained by averaging scores from the five dimensions, and the composite scores from a teacher's four lessons can then be aggregated into one overall teacher-level score. At the teacher-level, the choice to report composite lesson scores, the overall scores for each dimension, or individual subscale scores is dependent on the purpose and grain-size of the research questions. Generally, the overall scores for each of the five dimensions are reported. However, scores for each subscale within a dimension can be used to provide formative feedback directly to teachers or aggregated to inform professional development. Scores from multiple teachers over the course of several observations may also be aggregated to provide feedback at the district



level. Such data can indicate trends across teachers or schools for program evaluation purposes (e.g., to measure impact of professional development or changes in the curriculum).

The MQI as a Research Tool

Two of the original goals prompting the development of the MQI were (a) to provide more than a propositional link between teacher knowledge and classroom instruction and (b) to capture the mathematical aspects of instruction as distinct from pedagogical strategy. Next, we discuss examples of recent studies designed to meet these goals and explore other potential uses of the MQI.

Relating teacher knowledge and classroom instruction.

Hill and her colleagues at the University of Michigan (Hill et al., 2008) employed a mixed-methods approach to investigate how teachers' mathematical knowledge, as measured on paper-pencil MKT assessments, interacted with the mathematical quality of instruction. To facilitate the correlational study, MQI subscales were compressed into six separate dimensions (using an earlier version of the MQI). Teacher-level scores for each dimension and an overall lesson score were calculated by averaging scores across three videotaped lessons per teacher. Significant correlations were found between MKT and teachers' scores for the dimensions of *Total Errors*, *Language Errors*, and *Responding to Students Appropriately* (Spearman's $\rho = -.83$ and $-.80$ and $.65$, respectively). In addition to establishing strong positive correlations between MKT and MQI, qualitative analysis of both convergent (i.e., high MKT and high MQI) and divergent teacher cases revealed additional factors (i.e., curricular materials or teacher beliefs) that may mediate this relationship.

Building on this study, the MQI was used in a small-scale study to investigate the interrelationships amongst the mathematical quality of classroom instruction, curriculum, and teacher knowledge (Charalambous & Hill, 2012). Researchers aggregated overall scores on each MQI dimension across six videotaped lessons to categorize mathematical quality of instruction for each case-study teacher. Both MKT and the nature of the curriculum (e.g., reform-oriented or traditional) were positively related to teachers' abilities to use representations, provide explanations, and use precise language and notation as identified by the MQI.

Relating classroom instruction and students' mathematical achievement. MQI is one of five classroom observation protocols used in the larger Measures of

Effective Teaching (MET) project⁶ (Kane & Staiger, 2012). Preliminary findings indicate that the instrument does indeed measure qualities of instruction distinct from those assessed in non-subject specific observation tools (i.e., Classroom Assessment Scoring System [CLASS] or Framework for Teaching [FFT]). Specific studies within this project (see Appendix E) also reveal aspects of the MQI that are strongly or weakly correlated to teachers' value-added scores. Used at scale in this manner, the MQI can provide a vision of current mathematics instruction at a district or school level or indicate areas of instruction having the most impact on student achievement. For example, a comparison of scores across the five MQI dimensions indicated that the majority of mathematics lessons observed as part of the MET project were on-topic and relatively error free. However, the lessons were not necessarily mathematically rich, and students were given few opportunities to engage in sense-making activity (Kane & Staiger, 2012).

Summary of the MQI

The MQI instrument provides a reliable, quantifiable measure of the mathematical quality of instruction and is useful in a variety of settings. Two goals prompting the development of the MQI were (a) to provide more than a propositional link between teacher knowledge and classroom instruction and (b) to capture the mathematical aspects of instruction as distinct from pedagogical strategy. It can support large-scale studies attempting to establish correlations between classroom instruction and factors such as teachers' mathematical knowledge, professional development experiences, curricular materials, or student achievement. The level of detail provided within each subscale can also support qualitative analysis necessary to identify aspects of instruction that are especially enhanced or constrained by such factors.

Looking Across Classroom Observation Instruments

The classroom observation instruments reviewed herein (RTOP, IQA, MQI) have all been validated for use in mathematics education research. Each instrument can also be used to support mathematics teacher education, professional development, and program evaluation. Furthermore, each tool has features that may strengthen or limit its use, depending upon the specific contexts, resources, and research questions under investigation in a given study. In this section, we highlight features of the three tools that may help mathematics teacher

6 The MQI Lite, used in this comparative study, provides only overall scores for the five dimensions of the MQI, rather than scores for each subscale.

educators select a classroom observation tool appropriate for analyzing the specific aspect of instructional practice under investigation. We do this by considering the focus and contexts, as well as the features of each tool that may serve as affordances or constraints in relation to the goals, purposes, and resources of a given study.

Focus and Contexts

We propose that the purpose of a classroom observation instrument is to support knowledgeable raters' ability to notice the same aspects of instruction—ideally, aspects of instruction that impact students' learning of mathematics. Each tool reviewed herein helps the observer notice specific aspects of an instructional episode. What is noticed prompts the observer to make inferences about, and create meaning for, the instructional episode as a whole and specific events within that episode. With the multitude of events occurring throughout a mathematics lesson, each tool focuses the observer's attention in ways that elevate the importance of some events and reduce the importance of others.

The RTOP focuses the observer's attention on general features of reform-oriented instruction. Though the RTOP was developed to apply to mathematics and science, the reform-oriented instructional practices assessed by the RTOP rubrics are not inherently content-specific and may be applicable to other content-areas beyond mathematics and science. The large number of specific prompts within subscales makes the RTOP ideal for providing feedback to teachers regarding reform-oriented instruction. As the RTOP was designed to identify reform-oriented instructional practices, researchers or teacher educators should have some indication or expectation of the presence of these practices in order for the RTOP to be an appropriate tool. Contexts in which the RTOP may be ideal include assessments of (a) school-wide reform initiatives across multiple content areas, (b) preservice teachers and programs at the elementary level or across multiple secondary education content areas, or (c) professional development focused on the specific constructs identified by the RTOP (see Figure 1). For example, the RTOP would be a good choice to provide data for the research questions "Are preservice teachers in our program able to enact reform-oriented instruction?" (Jong, Pedulla, Reagan, Salomon-Fernandez, & Cochran-Smith, 2010) or "Did secondary mathematics and science teachers utilize reform-oriented instruction after experiencing such practices during university coursework?" (Adamson et al., 2003).

The IQA draws observers' attention to specific aspects of reform-oriented mathematics instruction, namely, cognitively challenging instructional tasks, task implementa-

tion, and discussion (including accountable talk). As the IQA was designed to identify specific reform-oriented instructional practices, some indication or expectation that these practices exist, are valued, or are intended to be developed over time should exist in order for the IQA to be appropriate to use. Contexts in which the IQA may be ideal include school-wide mathematics reform initiatives, preservice mathematics teachers and programs (at the elementary or secondary education level), and professional development, curriculum implementations, or large-scale assessments of reform-oriented mathematics teaching, *specifically focused* on the constructs identified by the IQA (e.g., cognitive demand and discussion). While similar to the RTOP in its focus on reform practices, the IQA attends to mathematics instruction and the use of cognitively challenging instructional tasks. Studies ideally suited for the IQA might investigate questions such as "Does secondary mathematics teachers' implementation of cognitively challenging tasks improve following their participation in task-centered professional development?" (e.g., Boston & Smith, 2009) or "How are elementary teachers using cognitively challenging tasks provided in their curriculum?" (e.g., Quint, Akey, Rappaport, & Willner, 2007).

The MQI assesses the rigor and richness of the mathematics throughout a lesson. In contrast to the other two tools, the MQI does not privilege reform-oriented instructional practices (though such practices may generate higher scores on some MQI rubrics, such as *Working with Students and Mathematics*, and *Common Core Aligned Student Practices*). The MQI is thus appropriate to evaluate students' opportunities to learn mathematics across a variety of instructional approaches, regardless of whether there is an expectation for reform-oriented instructional practices. MQI developers specify its use for grades K–9, perhaps because of the focus on mathematical content and the demands this places on raters. Contexts in which the MQI would be ideal include professional development initiatives or curriculum implementation, preservice mathematics teacher education programs at the middle or elementary level, or large-scale assessments of mathematics teaching, with a focus on the quality of mathematics during the instructional episode. Questions well-suited for the MQI include "What is the relationship between teachers' mathematical knowledge and the mathematical quality of instruction?" (e.g., Hill et al., 2008) or "How does the mathematical quality of instruction relate to students' mathematical achievement?" (e.g., Kane & Staiger, 2012).

In summary, the tools presented here overlap in some aspects but selectively attend to general reform-oriented practices (RTOP), specific reform-oriented practices in mathematics instruction (IQA), or the mathematical quality



of instruction across a variety of instructional approaches (MQI). The choice of an observation tool should be driven by the alignment between the research questions in a given study and the aspects of instruction made salient by that particular tool. We discuss next how features of each instrument can serve as affordances or constraints, depending on the focus and resources of a given study.

Features as Affordances or Constraints

As presented in each review, each classroom observation instrument has its own protocol for use and interpretation. We note that any feature is not inherently good or bad but becomes an affordance or constraint *in relation* to the goals and resources of a particular study. While our discussion references three specific tools, the four affordances and constraints discussed herein can apply to classroom observation tools more generally.

First, consider the requirement for live and/or videotaped observations. Live or videotaped observations can be used for the IQA or RTOP, and videotaped observations are necessary for the MQI. Live observations can be less invasive, which may facilitate participation and consent from schools, teachers, students, and parents. Live observations eliminate the need for video equipment or technology to collect, store, and share videos; however, they may require greater human resources, as at least one trained rater is necessary to observe and code each lesson. In live observations, the observer has access to the teacher, students, instructional materials, and other artifacts of the classroom and school. The rater can gain a holistic sense of classroom events (e.g., listening to conversations in multiple small groups; capturing participation, which may be difficult to identify in a video), where results obtained from videotape are influenced by the camera view of the lesson. However, live observations produce only written records and artifacts of the lesson, whereas video provides the ability to accumulate records of practice, score lessons independently or as a group, and replay instances in order to better understand key components of the lesson. Through technology, video can be shared in ways that enable raters to be in different physical locations (e.g., collaborations between researchers at different institutions).

Second, consider the requirement for rater training. The MQI and RTOP provide free, online rater training, while IQA training is only available face-to-face from the rubric developer. MQI online training can be completed individually, and raters require a certain level of mathematics knowledge (and mathematical knowledge for teaching) to successfully complete the certification. This system is thorough and produces high-quality raters but perhaps limits the pool of potential raters who can

achieve certification. RTOP training should be completed by the research group together to allow for discussion and consensus. Raters must have a general sense of reform-oriented practices; hence the potential pool of raters is larger than for the MQI. To use the IQA, research groups attend a training session provided by the developer. Raters develop the ability to classify tasks by level of cognitive demand (Stein & Smith, 1998), identify features of task implementation that serve to maintain or reduce cognitive demands (e.g., Henningsen & Stein, 1997), and identify specific features of discussions and accountable talk moves (e.g., Resnick & Hall &, 1998).

Third, consider how explicit each rubric is in providing descriptions of score levels. Rubrics can consist of detailed score levels (e.g., IQA and MQI) or more general, “sliding scale” score levels (e.g., the individual RTOP prompts), which create differences in using the instruments and interpreting the results. Using the RTOPs’ more general score levels, raters focus on the number of occurrences of each indicator. Results thus identify the presence or absence of reform-oriented practices and can serve to indicate which practices were lacking or not observed. The IQA (Appendix B) and the MQI (Appendix C) provide very descriptive score levels within each rubric or subscale. These score levels serve as explicit indicators of exactly what features of a construct were strong or need to be improved to achieve the next level (for raters and for providing feedback to motivate instructional improvements).

Fourth, consider the scale of the research project and the usability of the rubrics. For small-scale projects, the variety of RTOP prompts can provide rich descriptive data to foster conversations with teachers or evaluate teacher preparation programs or professional development workshops. The IQA and the MQI were designed to be used reliably by trained raters in large-scale studies. For the IQA, the limited number and narrow focus of the rubrics make the IQA useable regardless of the scale of the study. The systematic process for using the MQI (e.g., coding timed segments) contributes to its usability at scale. All of the instruments can provide detailed results and feedback at the subscale or rubric level. For any instrument, while summing or averaging across individual rubrics or subscales into a composite or overall score may be useful for research purposes (e.g., correlating observation results to student achievement data or tests of teacher content knowledge), collapsing subscales or constructs into a single score also reduces the level of detail and specificity for which the results can be reported and interpreted.

Figure 4 provides a summary of features of each instrument discussed herein. As presented, a given feature of a tool can serve as an affordance or constraint.

| | RTOP | IQA | MQI |
|--------------------|---|--|--|
| Focus | Reform-oriented instruction | Cognitively challenging tasks, implementation, and discussion | Mathematical quality of instruction |
| Contexts | Reform-oriented instruction: <ul style="list-style-type: none"> • Building-wide reform initiatives • Professional development initiatives • Preservice teacher education programs | Reform-oriented mathematics instruction: <ul style="list-style-type: none"> • Professional development initiatives • Curriculum implementation • Preservice mathematics teacher education programs • Large-scale assessments of reform-oriented mathematics teaching | Mathematics instruction: <ul style="list-style-type: none"> • Professional development initiatives • Curriculum implementation • Preservice mathematics teacher education programs • Large-scale assessments of mathematics teaching |
| Affordances | <ul style="list-style-type: none"> • General indicators of reform-oriented instruction • Can be used across content areas • Live or videotaped lessons • Teacher-level data • Validated in MET Study • Provides rich descriptive data for discussions with teachers • Can show change over time • Free on-line training | <ul style="list-style-type: none"> • Very specific focus; can make explicit connections to PD • Live or video-taped lessons, or student work • Can be used at scale; can provide school- or district-level data • Descriptive statistics reported on individual rubrics • Well-defined score levels, explicit about what is needed to achieve next score level • Inter-rater reliability • Validated in prior studies • Provides rich descriptive data for discussions with teachers • Can be used over time, across sites, or compared to prior research | <ul style="list-style-type: none"> • Not biased toward any type of instruction • Videotaped lesson observations • Can be used at scale; can provide school- or district-level data • Well-defined score levels, explicit about what is needed to achieve next score level • Inter-rater reliability • Validated in MET Study • Many rubrics: provides rich descriptive data for discussions with teachers • Can be used over time, across sites, or compared to prior research • Free online training |
| Constraints | <ul style="list-style-type: none"> • Not inherently mathematical • Many indicators—difficult to use at scale and to get exact-point agreement between raters • No descriptors for each score level; not explicit about what is needed to achieve next score level • Training videos do not depict K–12 instruction | <ul style="list-style-type: none"> • Limited focus • Bias toward reform-oriented mathematics teaching • Accessibility of training • Not appropriate for comparisons or evaluations when there is no expectation of reform-oriented mathematics instruction | <ul style="list-style-type: none"> • Limited to K–9 • Not intended for live observations • Raters need adequate MKT • Broader focus may make it harder to establish explicit connections to PD |

Figure 4. Summary of features of the RTOP, IQA, and MQI.



Using the Classroom Observation Instruments to Support Instructional Change

The classroom observation instruments reviewed herein are often used to evaluate teachers' instructional practices and provide feedback to inform research studies or interventions, teacher preparation courses or programs, or professional development efforts. While designed as research tools, each classroom observation instrument can also provide important information to support teachers' learning and instructional change by (1) serving as tools used in professional development and (2) providing a focus for formative assessment or self-evaluation of practice. Each of these uses will be discussed briefly, with specific reference to the classroom observation tools.

First, the classroom observation instrument themselves can be used as tools in professional development or teacher education. Used in this way, the tools could support teachers to notice the aspects of instruction central to each rubric, provide criteria for analyzing the aspects of instruction, and communicate a standard or develop a shared vision for practice. The RTOP has been used as a professional development tool for instructional planning, teaching, and reflecting on instruction (e.g., Ciancolo, Flory, & Atwell, 2006; Lawson et al., 2002; MacIsaac & Falconer, 2002). Lawson and his team (2002) used the RTOP to develop and frame summer institutes aimed at supporting college-level instructors to design and implement reform-oriented instruction. The IQA rubrics assess instructional practices (e.g., tasks, implementation, and/or discussion) that can be fostered through professional development (Boston & Smith, 2009, 2011). In fact, a variety of professional development materials currently exist that engage teachers in learning about and analyzing tasks, task implementation, and discussion (e.g., Stein & Smith, 1998; Stein & Smith, 2011; Stein, Smith, Henningesen, & Silver, 2009). Professional development or teacher education activities could be developed to incorporate the IQA rubrics explicitly (similarly for the RTOP and the MQI), where teachers might use the rubrics to analyze tasks, curricula, or videos of classroom instruction. The IQA has been used in this way to design a professional development workshop for middle-school principals. The goal of the workshop was to enable principals to identify high-quality tasks, implementation, and discussion during informal observations in mathematics classroom and to provide formative feedback to mathematics teachers based on these observations (Boston, 2011; Boston, Henrick, & Gibbons, 2014). Similarly, because the MQI is designed for use with video, it could easily be incorporated into professional development opportunities or

teacher education courses. In these settings, the MQI can raise teachers' awareness and understanding of critical components of quality mathematics instruction such as using precise mathematical language, linking representations, or focusing on patterns and generalizations.

Second, classroom observation tools can be used as tools for formative assessment or self-evaluations of practice. Teachers can use the tools to identify the presence and quality of specific practices and to provide concrete pathways for instructional change. Each tool has features that support formative assessment of practice. The RTOP identifies many specific items that allow for detailed feedback, and thus can be used to engage teachers in reflections or conversations about specific aspects of reform-oriented instruction. The IQA contains a small number of rubrics, but is designed with detailed score levels to indicate specific criteria for improving instruction within each rubric. For the MQI, the specificity provided in the rubrics, coupled with recent revisions to align the final dimension with CCSSM, can serve as a valuable learning tool as teachers begin to make sense of new standards. Clear examples and language can support teachers' reflection on how instruction (either their own or others') encompasses important mathematical practices. As a practice-based tool, the MQI can focus teachers' attention on key components of quality mathematics instruction such as using precise mathematical language, linking representations, or focusing on patterns and generalizations.

Summary and Conclusions

In this article, we have reviewed three classroom observation instruments to provide information about these tools and to present general considerations for selecting observation tools that would be useful to other mathematics teacher educators. The tools presented herein communicate and evaluate standards of practice in mathematics teaching and learning; specifically, of reform-oriented instruction (RTOP), the selection and implementation of cognitively challenging tasks and discussion (IQA), and/or the mathematical rigor and richness of a lesson (MQI). We propose that the selection of a classroom observation tool should depend upon the question under investigation and the focus (e.g., the aspects of instruction each tool helps you to notice) of a specific study, professional development project, or program evaluation. In this way, the tool generates evidence directly connected to the intervention, appropriate for assessing the impact or effectiveness of the intervention and establishing evidenced-based practice as stated in the call for *MTE* articles.

References

- Adamson, S., Banks, D., Burtch, M., Cox, F., Judson, E., Turley, J., . . . Lawson, A. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching*, 40, 939–957.
- American Association for the Advancement of Science. (1989). *Project 2061: Science for All Americans: A Project 2061 report on literacy goals in science, mathematics, and technology*. Washington, D.C.: Author.
- Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59, 389–407.
- Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Record*, 110, 608–645.
- Boston, M. D. (2011). *The Instructional Quality Assessment as a learning tool for school leaders*. Presentation at the annual meeting of the Association of Mathematics Teacher Educators, Irvine, CA.
- Boston, M. D. (2012). Assessing instructional quality in mathematics. *Elementary School Journal*, 113, 76–104.
- Boston, M. D., & Wilhelm, A. G. (in press). Middle school mathematics instruction in instructionally focused urban districts. Accepted by *Urban Education*, July, 2014.
- Boston, M. D., Henrick, E. C., & Gibbons, L. (2014). *Enabling principals to support high quality mathematics instruction*. Presentation at the National Conference of Supervisors of Mathematics, New Orleans, LA.
- Boston, M. D., & Smith, M. S. (2009). Transforming secondary mathematics teaching: Increasing the cognitive demands of instructional tasks used in teachers' classrooms. *Journal for Research in Mathematics Education*, 40, 119–156.
- Boston, M. D., & Smith, M. S. (2011). A "task-centric approach" to professional development: Enhancing and sustaining mathematics teachers' ability to implement cognitively challenging mathematical tasks. *ZDM: International Journal of Mathematics Teacher Education*, 43, 965–977. doi:10.1007/s11858-011-0353-2
- Boston, M. D., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction: The development of Instructional Quality Assessment Toolkit* (CSE Tech. Rep. No. 672). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Charalambous, C. Y., & Hill, H. C. (2012). Teacher knowledge, curriculum materials, and quality of instruction: Unpacking a complex relationship. *Journal of Curriculum Studies*, 44(4), 443–466.
- Charalambous, C. Y., Hill, H. C., & Mitchell, R. N. (2012). Two negatives don't always make a positive: Exploring how limitations in teacher knowledge and the curriculum contribute to instructional quality. *Journal of Curriculum Studies*, 44, 489–513.
- Cianciolo, J., Flory, L., & Atwell, J. (2006). Evaluating the use of inquiry-based activities: Do student and teacher behaviors really change? *Journal of College Science Teaching*, 36, 50–55.
- Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, 25, 119–142.
- Dunleavy, M., Dede, C., & Mitchell, R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of Science Education and Technology*, 18, 7–22.
- Henningsen, M. A., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education*, 28, 524–549.
- Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41, 56–64.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26, 430–511.
- Jackson, K., Garrison, A., Wilson, J., Gibbons, L., & Shahan, E. (2013). Exploring relationships between setting up complex tasks and opportunities to learn in concluding whole-class discussions in middle-grades mathematics instruction. *Journal for Research in Mathematics Education*, 44, 646–682.
- Jong, C., Pedulla, J., Reagan, E., Salomon-Fernandez, Y., & Cochran-Smith, M. (2010). Exploring the link between reformed teaching practices and pupil learning in elementary school mathematics. *School Science and Mathematics Journal*, 110, 309–326.
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (MET Project research paper, Bill and Melinda Gates Foundation). Available from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practitioner_Brief.pdf

- Lawson, A., Benford, R., Bloom, I., Carlson, M., Falconer, K., Hestenes, D., Judson, E., . . . Wyckoff, S. (2002). Evaluating college science and mathematics instruction. *Journal of College Science Teaching*, 31, 388–393.
- Learning Mathematics for Teaching Project (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14, 25–47.
- MacIsaac, D., & Falconer, K. (2002). Reforming physics instruction via the RTOP. *Physics Teacher*, 40, 479–485.
- Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions “at-scale.” *Educational Assessment*, 13, 267–300.
- Morrell, P., Wainwright, C., & Flick, L. (2004). Reform teaching strategies used by student teachers. *School Science and Mathematics Journal*, 104, 199–212.
- National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- National Governors Association Center for Best Practices & Council of Chief State School Officers (2010). *Common Core State Standards for Mathematics*. Washington, DC: Authors.
- National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.
- Quint, J. C., Akey, T. M., Rappaport, S., & Willner, C. J. (2007). *Instructional leadership, teaching quality, and student achievement: Suggestive evidence from three urban school districts*. New York, NY: MDRC. Retrieved October 28, 2014, from <http://www.mdrc.org/publications/470/execsum.html>.
- Resnick, L. B., & Hall, M. W. (1998). Learning organizations for sustainable education reform. *Daedalus*, 127, 89–118.
- Roehrig, G., & Kruse, R. (2005). The role of teachers’ beliefs and knowledge in the adoption of a reform-based curriculum. *School Science and Mathematics*, 105, 412–422.
- Sawada, D., Piburn, Judson, E., M., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2000). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics Journal*, 102, 245–253.
- Smith, M. S. (2014). Editorial: Tools as catalysts for practitioners’ thinking. *Mathematics Teacher Educator*, 3, 3–7.
- Smith, M. S., & Stein, M. K. (2011). *5 practices for orchestrating productive mathematics discussions*. Reston, VA: National Council of Teachers of Mathematics.
- Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in reform mathematics project. *Educational Research and Evaluation*, 2, 50–80.
- Stein, M. K., & Matsumura, L. C. (2008). Measuring instruction for teacher learning. In D. Gitomer (Ed.), *Measurement issues and the assessment of teacher quality* (pp. 179–205). Thousand Oaks, CA: Sage.
- Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics Teaching in the Middle School*, 3, 268–75.
- Stein, M. K., Grover, B. W., & Henningsen, M. A. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, 33, 455–488.
- Stein, M. K., Remillard, J., & Smith, M. S. (2007). How curriculum influences student learning. In F. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 319–369). Charlotte, NC: Information Age Publishing.
- Stein, M. K., Smith, M. S., Henningsen, M., & Silver, E. A. (2009). *Implementing standards-based mathematics instruction: A casebook for professional development*. New York, NY: Teachers College Press.
- Stigler, J. W., & Hiebert, J. (2004). Improving mathematics teaching. *Educational Leadership*, 61, 12–17.
- Sztajn, P., Wilson, P. H., Edgington, C., & Confrey, J. (2011). *Learning trajectories and key instructional practices*. Paper presented at the 33rd Annual Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education, Reno, NV.
- von Glasersfeld, E. (1989). Cognition, construction of knowledge, and teaching. *Synthese*, 80, 121–140.
- Wainwright, C., Morrell, P., Flick, L., & Schepige, A. (2004). Observations of reform teaching in undergraduate level mathematics and science courses. *School Science and Mathematics Journal*, 104, 322–335.
- Wilhelm, A. G. (2014). Mathematics teachers’ enactment of cognitively demanding tasks: Investigating links to teachers’ knowledge and conceptions. *Journal for Research in Mathematics Education*, 45, 636–674.

**Authors**

Melissa Boston, G14 Canevin, 600 Forbes Avenue,
Duquesne University, Pittsburgh, PA 15282;
bostonm@duq.edu

Jonathan Bostic, 529 Education Building, Bowling Green
State University, Bowling Green, OH
43403-0001;
bosticj@bgsu.edu

Kristin Lesseig, VUB 345, 14204 NE Salmon Creek Ave.,
Washington State University—Vancouver, Vancouver, WA
98686; kristin.lesseig@vancouver.wsu.edu

Milan Sherman, 211 Howard Hall, 2507 University Ave.,
Drake University, Des Moines, IA 50311;

milan.sherman@drake.edu

Appendix A: Sample Rubrics: Reformed Teaching Observation Protocol (RTOP)

IV. CONTENT

Propositional knowledge

| | | | | | | |
|-----|---|---|---|---|---|---|
| 6) | The lesson involved fundamental concepts of the subject. | 0 | 1 | 2 | 3 | 4 |
| 7) | The lesson promoted strongly coherent conceptual understanding. | 0 | 1 | 2 | 3 | 4 |
| 8) | The teacher had a solid grasp of the subject matter content inherent in the lesson. | 0 | 1 | 2 | 3 | 4 |
| 9) | Elements of abstraction (i.e., symbolic representations, theory building) were encouraged when it was important to do so. | 0 | 1 | 2 | 3 | 4 |
| 10) | Connections with other content disciplines and/or real world phenomena were explored and valued. | 0 | 1 | 2 | 3 | 4 |

Procedural Knowledge

| | | | | | | |
|-----|--|---|---|---|---|---|
| 11) | Students used a variety of means (models, drawings, graphs, concrete materials, manipulatives, etc.) to represent phenomena. | 0 | 1 | 2 | 3 | 4 |
| 12) | Students made predictions, estimations and/or hypotheses and devised means for testing them. | 0 | 1 | 2 | 3 | 4 |
| 13) | Students were actively engaged in thought-provoking activity that often involved the critical assessment of procedures. | 0 | 1 | 2 | 3 | 4 |
| 14) | Students were reflective about their learning. | 0 | 1 | 2 | 3 | 4 |
| 15) | Intellectual rigor, constructive criticism, and the challenging of ideas were valued. | 0 | 1 | 2 | 3 | 4 |

Note. A score of 0 indicates “never occurred.” A score of 4 indicates “very descriptive.”


[\(Return to page 155\)](#)

Appendix B: Sample Rubrics: Instructional Quality Assessment (IQA) Potential of the Task Rubric

| | |
|-----|---|
| 4 | <p>The task has the potential to engage students in exploring and understanding the nature of mathematical concepts, procedures, and/or relationships, such as:</p> <ul style="list-style-type: none"> • Doing mathematics: using complex and non-algorithmic thinking (i.e., there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instructions, or a worked-out example); OR • Procedures with connections: applying a broad general procedure that remains closely connected to mathematical concepts. <p>The task must explicitly prompt for evidence of students' reasoning and understanding.</p> <p>For example, the task MAY require students to:</p> <ul style="list-style-type: none"> • solve a genuine, challenging problem for which students' reasoning is evident in their work on the task; • develop an explanation for why formulas or procedures work; • identify patterns and form and justify generalizations based on these patterns; • make conjectures and support conclusions with mathematical evidence; • make explicit connections between representations, strategies, or mathematical concepts and procedures. • follow a prescribed procedure in order to explain/illustrate a mathematical concept, process, or relationship. |
| 3 | <p>The task has the potential to engage students in complex thinking or in creating meaning for mathematical concepts, procedures, and/or relationships. However, the task does not warrant a "4" because:</p> <ul style="list-style-type: none"> • the task does not explicitly prompt for evidence of students' reasoning and understanding. • students may be asked to engage in doing mathematics or procedures with connections, but the underlying mathematics in the task is not appropriate for the specific group of students (i.e., too easy or too hard to promote engagement with high-level cognitive demands); • students may need to identify patterns but are not pressed for generalizations or justification; • students may be asked to use multiple strategies or representations, but the task does not explicitly prompt students to develop connections between them; • students may be asked to make conjectures but are not asked to provide mathematical evidence or explanations to support conclusions |
| 2 | <p>The potential of the task is limited to engaging students in using a procedure that is either specifically called for or its use is evident based on prior instruction, experience, or placement of the task. There is little ambiguity about what needs to be done and how to do it. The task does not require students to make connections to the concepts or meaning underlying the procedure being used. Focus of the task appears to be on producing correct answers rather than developing mathematical understanding (e.g., applying a specific problem-solving strategy, practicing a computational algorithm).</p> <p>OR There is evidence that the mathematical content of the task is at least two grade-levels below the grade of the students in the class.</p> |
| 1 | <p>The potential of the task is limited to engaging students in memorizing or reproducing facts, rules, formulae, or definitions. The task does not require students to make connections to the concepts or meaning that underlie the facts, rules, formulae, or definitions being memorized or reproduced.</p> |
| 0 | <p>Students did not engage in a mathematical activity.</p> |
| N/A | |

[\(Return to page 158\)](#)

Appendix C: Sample Rubrics: Mathematical Quality of Instruction (MQI) 4-point

| Linking Between Representations | | | |
|--|---|--|---|
| <p>This code refers to teachers' and students' explicit linking and connections between different representations of a mathematical idea or procedure. To count, these links must occur across different representational "families" e.g., a linear graph and a table both capturing a linear relationship. So, two different representations that are both in the symbolic family (e.g., $1/4$ and 0.25) are not candidates for being linked.</p> <p>For Linking Between Representations to be scored above a Not Present:</p> <ul style="list-style-type: none"> At least one representation must be visually present The explicit linking between the two representations must be communicated out loud <p>For Linking Between Representations to be scored Mid or High, two conditions must be satisfied:</p> <ul style="list-style-type: none"> Both representations must be visually present The correspondence between the representations must be explicitly pointed out in a way that focuses on meaning (e.g., pointing to the numerator in $1/4$, then commenting that you can see that one in the figure, pointing to the four in the denominator, pointing to the four partitions in the whole. "You can see the 1 in the $1/4$ corresponds to the upper left-hand box, which is shaded, showing one piece out of four total pieces...") <div style="text-align: center;">  </div> <p>For geometry, we do not count shapes as a representation that can be linked—we consider those to be the "thing itself." However, links can be scored in geometry if the manipulation of geometric objects is linked to a computation, e.g., showing that two 45-degree angles can be combined to get a 90 degree angle and linking that to the symbolic representation $45 + 45 = 90$.</p> <p>Note: If links are made but underlying representation/idea is incorrect, do NOT count as linking between representations.</p> | | | |
| Not Present | Low | Mid | High |
| No linking occurs. Representations may be present, but no connections are actively made. | Links are present in a pro forma way; For example, the teacher may show the above figure and state that one quarter is one part out of four. These links will not be very explicit or detailed; both representations need not be present. | Links and connections have the features noted under High, but they occur as an isolated instance in the segment. | <p>Links and connections are present with extended, careful work characterized by one of the following features:</p> <ul style="list-style-type: none"> Explicitness about how two or more representations are <i>related</i> (e.g., pointing to specific areas of correspondence) OR Detail and elaboration about the relationship between two mathematical representations (e.g., noting meta-features; providing information about under what conditions the relationship occurs; discussing implications of relationship) <p>These links will be a characterizing feature of the segment, in that they may in fact be the focus of instruction. They need not take up the majority or even a significant portion of the segment; however, they will offer significant insight into the mathematical material.</p> |

| Overall Richness of the Mathematics | | | |
|---|--|---|---|
| This code captures the depth of the mathematics offered to students. | | | |
| Note: This is an overall code for each segment. It is not an average of the codes in this dimension, but an overall estimate of richness. | | | |
| Not Present | Low | Mid | High |
| <p>Elements of richness are present but are all incorrect</p> <p>OR</p> <p>Elements of rich mathematics are not present.</p> | <p>Elements of rich mathematics are minimally present.</p> <p>Note that there may be isolated Mid scores in the codes of this dimension.</p> | <p>Elements of rich mathematics are more than minimally present but the overall richness of the segment does not rise to the level of a High.</p> <p>For example, a segment may be characterized by some Mid scores in the codes of this dimension or by an isolated High along with substantial procedural focus, etc.</p> | <p>Elements of rich mathematics are present, and either:</p> <p>a) There is a combination of elements that together saturate the segment with rich mathematics either through meaning or mathematical practices.</p> <p>OR</p> <p>b) There is truly outstanding performance in one or more of the elements.</p> |
| Scoring Help - Overall Richness of the Mathematics | | | |
| <p>In scoring Overall Richness, we assign a score of Not Present when there are no elements of richness present in the segment, or the components of richness that are present are all incorrect. For this code, we do not consider middling density of Mathematical Language to be an element of richness. That is, a segment could get a score of Low or Mid for Mathematical Language and still get a score of Not Present for Overall Richness.</p> | | | |

MQI 4-point ©2014 Learning Mathematics for Teaching/Heather Hill.

[\(Return to page 161\)](#)

Appendix D: Additional References and Resources

- This article is based on a session at the AMTE 2014 Annual Meeting. The AMTE session PowerPoint can be downloaded at: <http://padlet.com/wall/amte2014>
- Reformed Teaching Observation Protocol (RTOP)
- Training Site:
http://physicsed.buffalostate.edu/AZTEC/RTOP/RTOP_full/index.htm
- Entire Protocol:
http://physicsed.buffalostate.edu/AZTEC/RTOP/RTOP_full/PDF/RTOPform_IN001.pdf
- Research using the RTOP: Adamson et al., 2003; Dunleavy, Dede, & Mitchell, 2009; Jong, Pedulla, Reagan, Salomon-Fernandez, & Cochran-Smith, 2010; Roehrig & Kruse, 2005.
- Adaptations of RTOP: IOP in Ciancolo, Flory, & Atwell, 2006; see OTOP in Morrell, Wainwright, & Flick, 2004; Wainwright, Morrell, Flick, & Schepige, 2004
- IQA rubrics are available for viewing at:
http://peabody.vanderbilt.edu/docs/pdf/tl/IQA_RaterPacket_LessonObservations_Fall_12.pdf
- For information on IQA training, contact Melissa Boston at: bostonm@duq.edu
- Research using the IQA: Boston, 2012; Boston & Smith 2009, 2011; Boston & Wilhelm, in press; Quint, Akey, Rappaport, & Willner, 2007; MIST Project:
http://peabody.vanderbilt.edu/departments/tl/teaching_and_learning_research/mist/index.php
- Learning Mathematics for Teaching (LMT) Project:
<http://www.sitemaker.umich.edu/lmt/home>
- Mathematical Quality of Instruction (MQI) Training Site:
http://isites.harvard.edu/icb/icb.do?keyword=mqi_training
- Information on MQI studies based on the MET data is available at:
<http://www.gse.harvard.edu/ncte/projects/core/default.php#.U6hZNqgozdA>

[\(Return to page 155\)](#)