The Mathematics Educator
2022 Vol. 30, No. 2, 33–63

Does the Choice of Observation Instrument Matter?

Jennifer M. Lewis, S. Asli Özgün-Koca, Lenuel Hernandez, Christopher Nazelli, and Kate R. French

Does the choice of observation instrument make a difference in the feedback and ratings that teachers receive? This study explores how lessons are rated differentially across various observation instruments. To investigate this question, ten randomly selected mathematics lessons were rated using six different observation instruments. Overall scores varied little across instruments. Our analyses indicate that differences in scores can be attributed to what we call instrumental occlusion, instrumental emphasis, and element density. This article concludes with implications for the selection and use of observation instruments in school settings.

Among all in-school factors, teachers' instructional practice has the biggest impact on student achievement (Chetty et al., 2014; Darling-Hammond et al., 2009). In order to improve teachers' instructional practice, observers have watched instruction and provided feedback to teachers, formally and informally. Increasingly, these observations are also used to make high-stakes decisions regarding teacher assignments, promotions, demotions, salary, etc. (Hull, 2013; Millman, 1981; Popham, 2013). The conversations surrounding an observation

Jennifer Lewis *is an associate professor of mathematics education and teacher education at Wayne State University. She is especially interested in how teachers learn mathematics in the context of their daily work.*

S. Asli Özgün-Koca *is a professor of mathematics education at Wayne State University. Her primary research focuses on the use of technology in mathematics education and mathematics teacher education.*

Lenuel Hernandez *is the mathematics coordinator for the Academic Success Center at Wayne State University. His primary research focuses on help seeking of higher education students and mathematical thinking study habits.*

Christopher Nazelli *is an associate professor of teaching in Wayne State University's Department of Mathematics. His research focus is mathematics teacher education.*

Kate R. French *is an assistant professor and chair of teacher education at Madonna University. Her work focuses on how policies, processes, and sociocultural context shape teacher learning and practice.*

can also create a shared language for discussing instruction, hold teachers accountable for what they do, and constitute the basis for critical colleagueship (Lord, 1994). The development of observation instruments has grown apace with the expanded use of observations for multiple purposes.

Observation instruments, as currently conceived, were first developed in the 1970s, when efforts in educational improvement shifted from a focus on curriculum to a focus on instruction (Kersten & Israel, 2005). Absent a consensus around the definition of effective teaching, especially in mathematics, the foci of instruments varied, although most existing observation instruments seemed valid to stakeholders (Goe et al., 2008). As notions of effective teaching evolved, new observation and evaluation tools followed (Fenstermacher & Richardson, 2005). The press for accountability in schools led to an increase in the use of observation instruments as part of an expanded effort to evaluate teachers and motivate better instruction (No Child Left Behind [NCLB], 2002). Information gathered from observations can identify strengths and challenges in instruction which can then be leveraged to make teaching and learning more effective. Observation instruments have the potential to help observers notice aspects of instruction and provide feedback to the teacher for improvement and reflection. Different observation instruments emphasize different aspects of teaching—some are designed to appraise general pedagogy while others are developed around content-specific pedagogies, and within those two categories emphases can vary as well (Blazar et al., 2017).

When teacher observations contribute to high-stakes employment decisions, the concern for quality and direction of teacher observation ratings is heightened (Chetty et al., 2014). Observation instruments abound and may produce different ratings or emphasize different facets of instruction. Hence, we pose the following question: Does the choice of observation instrument affect ratings? To investigate this research question, we designed a study to explore whether mathematics lessons are rated differentially across different observation instruments.

General and Content-Specific Observation Instruments

Creating an instrument to capture *all* the elements of classroom instruction is virtually impossible (Goe et al., 2008; Schoenfeld, 2013). One is reminded of Umberto Eco's essay, "On the Impossibility of Drawing a Map of the Empire on a Scale of 1:1" (1994), in which the map of a place is so detailed that the map becomes isomorphic with the city itself. Of necessity, each instrument emphasizes certain elements of classroom instruction and is therefore more detailed around those elements, while leaving other elements less specified.

Both general and content-specific instruments are used to evaluate mathematics teaching (Blazar et al., 2017). General instruments are often easier and more practical to implement, oversee, and manage for raters from various backgrounds. They are helpful in assessing aspects of instruction such as classroom management, student engagement, and the culture of learning. Content-specific instruments are designed to capture those aspects of subject-matter instruction relevant only to the discipline (Goe et al., 2008; Hill & Grossman, 2013; Learning Mathematics for Teaching Project, 2011). Content-specific instruments can be used by raters with appropriate content knowledge to produce targeted and actionable feedback (Hill & Grossman, 2013). However, content-specific instruments may produce inconsistent evaluations reflecting rater biases if evaluators have limited subject-matter knowledge (Goe et al., 2008).

Theoretical Framework

Observation instruments reflect conceptions of effective teaching. Implicit in observation instruments for appraising teaching is the notion that worthwhile and observable teacher actions lead to student learning and should therefore be noticed, valued, and strengthened (Chetty et al., 2014). To compare ratings on different observation instruments, we rely on two different conceptualizations of teaching. Our approach to understanding mathematics-specific instruction is framed by

Ball et al.'s (2008) theory of "Mathematical Knowledge for Teaching." By examining the real work of teaching mathematics, Ball et al. (2008) were able to create a practice-based model of the knowledge needed to carry out mathematics instruction. Their analysis of the tasks and problems that arose repeatedly in instruction, and the knowledge that teachers drew on to carry out and address them served to delineate broad domains of content and pedagogical knowledge.

The second conceptualization of teaching that underlies our study comes from Maulana et al. (2017), who articulated a collection of empirically-based teaching behaviors[1] that lead to student achievement. These behaviors, relevant to the teaching of all subjects, include "creating a safe and stimulating learning climate, efficient classroom management, providing clear instruction, activating learning, adaptive teaching and teaching learning strategies" (p. 473).

We chose to use both the Mathematical Knowledge for Teaching framework (Ball et al., 2008) and the Model of Effective Teaching Behaviour (Maulana et al., 2017) because each model represents essential features of teacher action in instruction, both the mathematics-specific and the general. Table 1 displays the core ideas from these two conceptualizations of teaching that form the basis for our theoretical framework.

Table 1
Core Ideas about Instruction

Framework and *Components*	Descriptions
Mathematical Knowledge for Teaching (Ball et al., 2008; Hill et al., 2007)	Mathematical Knowledge for Teaching (MKT) is a conceptualization of the knowledge needed to teach mathematics. It derives from a study of the recurrent behaviors and strategies that teachers use as they teach mathematics. MKT encapsulates both content and pedagogical knowledge. The components of

[1] Maulana et al. (2017) use the term "teaching behaviours," so we have preserved that term in the title and Americanized the spelling elsewhere.

	MKT that are relevant to our work are described below.
Common Content Knowledge	The mathematical knowledge that teachers seek to develop in their students. This knowledge includes the facts, skills, and conceptual understandings employed by all users of mathematics.
Specialized Content Knowledge	The knowledge used only in the *teaching* of mathematics. This knowledge is used, for example, when responding to student thinking or facilitating mathematical discussions.
Knowledge of Content and Students	The blend of content knowledge and the knowledge of how students learn particular topics and the typical misconceptions that arise. Teachers may draw on this knowledge when diagnosing the cause of student errors, for example.
Knowledge of Content and Teaching	The knowledge of how best to represent mathematical ideas through examples, visuals, and discussions during instruction.
Framework: Model of Effective Teaching Behaviour (Maulana et al., 2017)	This theoretical model notes visible teaching behavior, categorized as "effective when it has a significant influence on student outcomes such as academic engagement" (p. 473) in a classroom setting, based on evidence from the literature. The components of effective teaching are described below.
Clarity of instruction	Provides clear, coherent instruction for the duration of lessons; provides students with greater capacities to learn and perform. Clear alignment of lesson presentations, activities, and group work are also important components of instructional clarity.
Adaptive teaching	Recognizes the varied needs and characteristics of students and responds to those needs through

	varied instructional approaches and strategies.
Activating learning	Engages students in active and challenging forms of learning that build on prior knowledge and are relevant to lesson content. Teacher acts as a facilitator of learning. As the quality of teacher-student interactions increases, so does student performance.
Teaching learning strategies	Incorporates opportunities for metacognition (i.e. ability for students to evaluate their learning processes and their understanding of content and skills). Teacher scaffolds lessons to appropriately build from existing skills toward desired skills.
Safe and stimulating learning climate	Creates an environment that is optimal for learning. This environment requires an atmosphere that is supportive, comfortable, encouraging, and instills a mutual respect amongst peers and between student and teacher.
Efficient classroom management	Uses time judiciously, ensures lesson continuity and quality, minimizes distractions, and addresses student misbehavior quickly and constructively.

Note. The titles of the two frameworks are bolded in the left column while the frameworks' components are italicized.

These core ideas about instruction were the basis for nine constructs developed for this study to compare the use of different observation instruments in evaluating mathematics lessons. The Mathematical Knowledge for Teaching framework was most useful in guiding the creation of constructs pertaining to the content knowledge teachers need to teach mathematics, while the Model of Effective Teaching Behaviour was most useful in developing constructs of general pedagogy. In light of these important contributions, we felt that it was necessary to incorporate both models into our theoretical framework. The development of these constructs is described in the Data and

Methods section below, with much more detail provided in our coding manual.[2]

Data and Methods

To investigate our research question, the research team reviewed ten randomly selected videos of mathematics lessons featuring teachers of grades 4 through 8. The teachers in these videos had been participants in a professional development program unrelated to this research project and geographically distant from the research team. Participating teachers were videotaped at multiple stages of their professional development program, creating a large video dataset. The dataset was known to the research team and was both convenient and distant. A random selection of ten videos was requested as a way to simulate the kind of teacher observations that many raters face in practice: a wide range of unrelated lessons taught by teachers of varying levels of experience and expertise. To obtain this set of lessons, the primary investigator of the professional development program was asked to share ten random videos, without knowing the research purposes of this study.

Instrument Selection and Construct Development

To study the variation in lesson ratings across different instruments, we purposefully chose six teacher observation instruments to rate ten randomly chosen mathematics lessons. Two of our instruments were general and four were mathematics-specific. To be included in this study, the instruments needed to be widely used in practice, derived from research, and familiar to the research team.

The research team selected two general observation rubrics: Danielson's *Framework for Teaching* (Danielson, 2013), and Marzano's *Teacher Evaluation Model* (Marzano, 2013). Both were chosen because of their ubiquity as teacher evaluation tools in states across the nation; at the time of the study both were in

[2] For more information on our coding procedures, please contact the authors.

use by many school districts surrounding our university, including the largest public district. The four mathematics-specific instruments used in this study were *Mathematics Scan* (M-Scan) (Berry III et al., 2010), The *Mathematical Quality of Instruction* (MQI) (Hill et al., 2008), *Reformed Teaching Observation Protocol* (RTOP) (Sawada & Piburn, 2000), and *Teaching for Robust Understanding of Mathematics* (TRU Math; Schoenfeld, 2013).

Collectively, the seven members of the research team had training and extensive experience using the six instruments. One member of the research team was a developer and trainer for MQI and had used the instrument in multiple research projects; she trained others on the instrument as well. Two research team members had used Danielson's *Framework for Teaching* in large-scale research projects previously and had been trained in that context. Three members of the research team had been trained at the university to use RTOP in college-level science and mathematics courses. One member of the research team had been trained to use Marzano's *Teacher Evaluation Model* for a school district; another had extensive experience using TRU in an evaluation of a professional development program. Some research team members had been trained to use more than one of the instruments.

All members of the research group have backgrounds in mathematics education and experience as teachers; four hold advanced degrees in mathematics. Each research team member scored all ten lessons using the two instruments that they had been trained to use, so that each lesson was scored using all six instruments. Researchers first scored all ten videos and produced scores independently on a single instrument. Then pairs of researchers using the same instrument reconciled their ratings where possible but preserved both sets of ratings to show differences if differences remained.

Video analysis followed Erickson's method, specifically what he refers to as the "manifest content approach" of video analysis (Erickson, 2006, p. 186). The researchers used Erickson's protocol for microanalysis of videotaped classroom data by taking each lesson as a unit of analysis and following Erickson's steps:

1. Viewing the whole lesson in its entirety
2. Identifying major constituent parts of the lesson
3. Identifying aspects of organization within major parts of the lesson
4. Focusing on actions of individuals within the lesson
5. Analyzing comparative instances across the lesson

Throughout the process, researchers were guided by Erickson's recommendation to combine attention to the subject-matter substance and the social interactions in each lesson.

In addition to the research team, two school principals were invited to rate two randomly selected videos from the set of ten using a general observation instrument. Neither principal had specialized knowledge or experience in teaching mathematics; both had extensive experience using the Danielson Framework for Teaching for teacher evaluation, which is the instrument they used here. We conducted cognitive interviews (Desimone & Le Floch, 2004) with these two principals afterwards to understand possible contrasts between researchers' perspectives and practitioners' use of the instruments, as well as how these two principal raters outside the research team may view lessons differently. There was no noticeable difference in the ratings created by the two principals as compared with the ratings of the research team members using the Danielson Framework for Teaching: the principals and the researchers gave the same overall ratings.

Comparing Ratings Across Instruments

In order to compare ratings of one lesson across the six different instruments, the research team first developed nine constructs based on the two conceptualizations of teaching described earlier—the Mathematical Knowledge for Teaching framework (Ball et al., 2008) and the Model of Effective Teaching Behaviour (Maulana et al., 2017). Based on the research team's collective work on a content analysis of these two conceptualizations, these nine constructs were developed, grouping like elements from the different instruments into categories that aligned with the two conceptualizations of

teaching. We attempted to link every element of each instrument to one of the nine constructs, so that ratings across the different instruments could be compared for a single lesson. To give the reader a sense for this comparison tool, the working definition of each construct and its connection to the Mathematical Knowledge for Teaching (MKT) and Model of Effective Teaching Behaviour (METB) components is given in Table 2.

Table 2

Description of Constructs for Comparing Ratings

Working Definition	Component
1. Mathematical accuracy: The teacher has a strong command of mathematics content and makes no mathematical errors.	**MKT**: Common Content Knowledge, Specialized Content Knowledge
2. Mathematical quality of task: The teacher designs and implements a lesson that involves "significant and worthwhile mathematics" (National Council of Teachers of Mathematics, 1991), assigns tasks that go beyond application of procedures, and provides opportunities to develop conceptual understanding.	**MKT**: Knowledge of Content and Students, Knowledge of Content and Teaching
3. Mathematical practices: The teacher engages student in mathematics through 1) representations and tools; 2) justification and explanation; 3) problem solving; and 4) connections and applications. Students make predictions with explanations, and give answers supported by explanations.	**MKT**: Knowledge of Content and Teaching
4. Lesson design, coherence, and implementation: Components of the lesson are conceptually coherent and build upon each other.	**MKT**: Knowledge of Content and Teaching**; METB**: Clarity of Instruction
5. Teacher assessment of student knowledge: The teacher uses multiple methods to gauge each student's knowledge. The teacher has concrete evidence of student knowledge and	**METB**: Adaptive Teaching

Working Definition	Component
understanding, and uses this in instruction.	
6. Students' active participation and direction: Students are actively engaged in the lesson. Student productions shape the outcomes of how class time is spent, and what/how material is discussed. Students are doing the thinking in the class.	**METB**: Activating Learning, Teaching Learning Strategies
7. Teacher's responsiveness to students: The teacher recognizes students' academic needs and addresses them individually or within the whole group.	**METB**: Adaptive Teaching, Teaching Learning Strategies
8. Communication, respect, and rapport: Communication (teacher-student or student-student) in the classroom is conducted with respect and assignment of competence. Teacher has good rapport with students.	**METB**: Safe and Stimulating Learning Climate
9. Management: The teacher uses materials, such as manipulatives and instructional tasks, and behaviors, such as classroom routines, procedures, and expectations, that support learning.	**MTEB**: Efficient Classroom Management

Pairs of researchers mapped the elements of each instrument to the constructs with which they aligned, recorded the rationales for their mappings in analytic memos, and reconciled differences through discussion. Table 3 shows the number of elements from each of the six observation instruments in these constructs. Note that some elements in an instrument were counted under multiple constructs.

Table 3
Number of Elements from Each Instrument Associated with the Nine Constructs

	CONSTRUCT 1	CONTRUCT 2	CONSTRUCT 3	CONSTRUCT 4	CONSTRUCT 5	CONSTRUCT 6	CONSTRUCT 7	CONSTRUCT 8	CONSTRUCT 9	TOTAL
	Mathematical accuracy	Mathematical quality of task	Mathematical practices	Lesson design, coherence, & implementation	Teacher assessment of student knowledge	Student active participation & and direction	Teacher responsiveness to students	Communication, respect, & rapport	Classroom Management	
Marzano	0	2	9	15	3	10	1	10	12	62
Danielson	2	5	2	6	3	2	4	2	4	30
MSCAN	1	2	11	6	2	5	3	1	0	31
MQI	3	10	7	3	1	6	3	0	0	33
RTOP	1	5	8	5	1	8	5	6	0	39
TRU	1	3	5	2	3	7	14	0	2	37

To help illustrate how elements from the different instruments fit into the nine constructs, let us consider Element 11 of the Marzano instrument, one of nine elements from Marzano placed in Construct 3: Mathematical Practices. It states, "The teacher asks questions or engages students in activities that require elaborative inference that go beyond what was explicitly taught" (Marzano, 2013, p. 13). As evidence of this element, observers consider if "students *provide explanations and 'proofs'* for inferences" for their work (Marzano, 2013, p. 13; emphasis ours). The researchers saw this as a strong indication of justification and explanation, as defined in Construct 3.

To further explicate the construct development and assignment of corresponding elements, we offer another example here. Eleven elements of the M-Scan instrument (called "dimensions" in that instrument) were classified as addressing the presence of mathematical practices, Construct 3. Of the eleven elements, "Explanation and Justification Dimension" is strongly aligned with this construct as it directs the observer to notice if the teacher's actions encourage students to *justify their mathematical claims* (Berry III et al., 2010, p. 21).

Several of the instruments include indicators as evidence for its elements. Although the element, as a whole, may strongly align with one construct, the indicators may have an association with other constructs. For example, Danielson's element (or, in the vernacular of the instrument, sub-domain) 3B: Using Questioning and Discussion Techniques prompts the observer to look for the following indicators:

- Questions of high cognitive challenge, formulated by both students and teacher
- Questions with multiple correct answers or multiple approaches even when there is a single correct response
- Effective use of student responses and ideas
- Discussion, with the teacher stepping out of the central, mediating role
- Focus on the reasoning exhibited by students in discussion, both in give-and-take with the teacher and with their classmates

- High levels of student participation in discussion (Danielson, 2013, p. 65).

This single element aligned with five of our constructs. The researchers saw alignment with the fifth indicator and Construct 3: Mathematical Practices and alignment with the sixth indicator and Construct 6: Student Active Participation and Direction. The research team also saw a weak alignment with several indicators and Constructs 2, 5, and 7.

It is important to note that, for some instruments, particular elements were not associated with any of the nine constructs. For example, Danielson's element 1d: Demonstrating Knowledge of Resources measures the availability and variety of resources for students to support, reinforce, and extend learning as well as the availability of professional learning resources for the teacher, which is not associated with any of the nine constructs. It is also the case that the instruments did not necessarily contain elements in all 9 constructs. As shown in Table 3, the Marzano instrument had no elements associated with Construct 1: Mathematical Accuracy, and M-Scan, MQI, and RTOP had no elements associated with Construct 9: Classroom Management.

Because each observation instrument had its own scoring scale, our research team created composite scores using weighted ratings from constructs. To reach these overall ratings of high, medium, and low, researchers computed the weighted average for all elements in each construct. For example, the rating scale for the Danielson instrument includes four levels: 1 (Unsatisfactory), 2 (Basic), 3 (Proficient), and 4 (Distinguished). If the weighted average on one of our constructs fell in the interval from 1 and up to 2, the researchers assigned an overall rating of "Low." Weighted averages that fell in the intervals beginning at 2 and including 3 were assigned the overall rating of "Medium;" and any score above 3 was assigned the overall rating of "High." The research team used a similar weighted average conversion system for each instrument.[3]

[3] For further information on the calculation of weighted scores and the coding manual, please contact the authors.

Once the ten videotapes were coded by pairs of researchers, the full team inspected the table of ratings across all videos to note patterns and themes across and within the videotaped lessons.

Findings

Tables 4 and 5 below show ratings across all six instruments for two different lessons—which we chose as representative examples of the ten lessons that were analyzed. In total ten such tables were created, displaying the raters' scores for the ten lessons that were analyzed. Here we describe the results of our ratings across all ten lessons.

Looking across all ten lessons, we see that the choice of observation instrument made little difference on the scores for most of the videotaped lessons we viewed. Roughly 86% of the time (6 out of 7 instances), lessons rated low using a general observation instrument were also rated low with a mathematics-specific instrument. These results suggest that ratings are typically consistent across different instruments. In the rare event when instruments do produce different ratings, differences matter. We found that these differences fell into three categories: instrumental occlusion, instrumental emphasis, and element density. We explain these categories here.

Instrumental Occlusion

The first type of difference we attribute to what we call "Instrumental Occlusion." This is when different scores are present between content-specific items and non-content-specific items within instruments. We find that the content-specific items were crucial in differentiating between lessons that might look engaging but are mathematically weak and/or lessons that might be cognitively demanding mathematically but are weak in general pedagogical areas. We consider the lesson which we refer to as the "Candy Bar Lesson," with ratings shown in Table 4, as an example.

It is important to note that the ratings by both researchers who scored the lesson with each instrument are recorded in separate lines. Rather than reconciling these differences between the converted low-medium-high ratings through conversations, all ratings are preserved with an explanatory note. The differences between the two raters' scores were on the small side— for example, between low and medium or medium and high scores.

The Candy Bar Lesson involved students engaging with a hands-on task in small groups. Students measured the dimensions of a miniature-size candy bar and were then asked to create a larger, proportional drawing of its full-size version. The scores in the constructs in the last three columns indicate that classroom management, climate, and communication were strong. Scores in the first three content-specific constructs, however, indicate that the lesson was weak mathematically. The degree to which a particular instrument focuses the rater's eye on content matters—see, for example, the differential focus on Mathematical Accuracy between the Marzano instrument and the TRU. Thus, in the Candy Bar Lesson, what seemed like an engaging, hands-on lesson with lots of mathematical figuring in the eyes of an observer with little expertise in mathematics was actually fairly weak through the lens of more mathematically attuned instruments: despite all the student engagement, there was little mathematical thinking for students to do. The teacher supplied almost all of the mathematical calculation and reasoning in this lesson, and although students looked to be actively engaged in creating drawings, the teacher was directing them what to write and where to write it. The scores for classroom management and mathematical content diverged across instruments for this lesson. The divergence between scores on the mathematical constructs (in the first three columns in our tables) and scores on the general pedagogy constructs (the last two columns in our tables) appeared in multiple lessons and across all six instruments, but especially Danielson, Marzano, and the RTOP instruments.

Table 4
Ratings for the Candy Bar Lesson

Generic Instruments	Mathematical accuracy	Mathematical quality of task	Mathematical practices	Lesson design, coherence, & implementation	Teacher assessment of student knowledge	Students' active participation and direction	Teacher's responsiveness to students	Communication, respect and rapport	Management
Danielson	LOW	LOW	LOW	LOW	LOW	LOW	LOW	MED	MED
	LOW	LOW	MED	LOW	MED	MED	LOW	MED	MED
Marzano	N/A	LOW	LOW	MED	MED	LOW	MED	MED	HIGH
	N/A	LOW	LOW	MED	LOW	LOW	MED	MED	MED
Subject Specific									
RTOP	LOW	LOW	LOW	LOW	LOW	LOW	LOW	MED	N/A
	LOW	LOW	LOW	LOW	LOW	LOW	LOW	MED	N/A
MSCAN	MED	LOW	LOW	MED	MED	MED	MED	MED	N/A
	MED	LOW	LOW	MED	LOW	MED	MED	MED	N/A
MQI	MED	LOW	LOW	MED	LOW	MED	LOW	N/A	N/A
	MED	MED	LOW	MED	LOW	MED	LOW	N/A	N/A
TRU	LOW	MED	MED	MED	MED	MED	MED	N/A	MED
	LOW	MED	MED	MED	MED	MED	MED	N/A	HIGH

Instrumental Emphasis

The second type of difference we attribute to what we term "Instrumental Emphasis," meaning that instruments have different emphases. This appears in several different ways. Some differences occur between instruments. Content-specific instruments can produce different ratings than the general pedagogy instruments. Among the mathematics-specific instruments are varying emphases as well. These differences include the following:

- Instruments vary in specification about teaching and learning of mathematics. Some are highly detailed about mathematical practices; some are more general.
- Instruments vary in the degree of specification of mathematical content. For example, in Mathematical Accuracy, MQI provides more specification than TRU.
- Instruments highlight mathematical practices differentially. For example, RTOP has more elements that focus on the practice subcategory *justification and explanation*, while M-Scan has more elements that focus on the practice subcategory *representation and tools*.
- Instruments may be more sensitive to certain types of mathematics lessons than others, and lesson types vary. For example, an introductory lesson may have a different architecture than an exploratory lesson, a lesson may feature more hands-on work, or may vary from whole-class discussion of earlier material. These differences became apparent when we examined discrepancies in the ratings.

Table 5 shows the scores from raters for all six instruments for a lesson in which students study the relationship of head circumference and height, so we refer to it as the "Head/Height Lesson." We consider this lesson because it has the interesting feature of divergent scores across types of instrument: low to medium scores on the general observation instruments but medium to high scores on the mathematics-specific instruments, shown in Table 5. In fact, it is the one lesson out of the ten that

we analyzed where the scores diverge markedly across instruments. The lesson opened with students collecting their head and height measurements and recording them in a table. There were students standing up and moving around the classroom, and it seemed that most of the movement and discussions were connected to the mathematical focus of the lesson. Students then entered their measurements on a shared table (projected by document camera) including the ratio of height to head circumference, shown in Figure 1 below. At different points in the lesson, the teacher stopped to ask students to look for relationships within the data.

Figure 1

Height and Head Circumference Table, Projected in Class

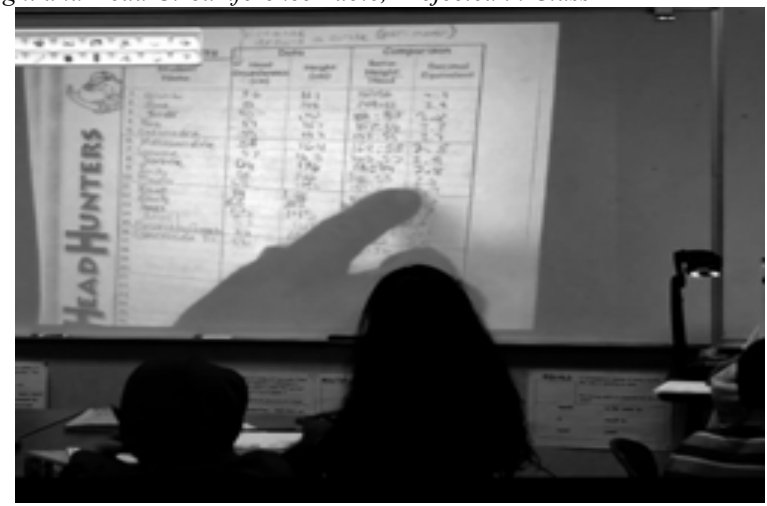

Table 5
Ratings for the Head/Height Lesson

Generic Instruments	Mathematical accuracy	Mathematical quality of task	Mathematical practices	Lesson design, coherence, & implementation	Teacher assessment of student knowledge	Students' active participation and direction	Teacher's responsiveness to students	Communication, respect and rapport	Management
Danielson	MED	MED	MED	MED	MED	MED	MED	MED	MED
	MED	MED	MED	MED	MED	MED	MED	MED	MED
Marzano	N/A	LOW	LOW	LOW	LOW	LOW	MED	MED	MED
	N/A	MED	MED	LOW	MED	MED	MED	LOW	MED
Subject Specific									
RTOP	MED	HIGH	MED	MED	LOW	MED	HIGH	MED	N/A
	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	MED	N/A
MSCAN	MED	MED	MED	MED	MED	LOW	MED	LOW	N/A
	HIGH	LOW	LOW	HIGH	MED	MED	MED	LOW	N/A
MQI	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	N/A
	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	N/A
TRU	MED	MED	LOW	MED	LOW	LOW	LOW	N/A	MED
	HIGH	MED	HIGH	HIGH	MED	MED	MED	N/A	MED

This example illustrates how instruments can vary in specification, substance, how mathematical practices are highlighted, and how they vary in terms of sensitivity toward certain types of mathematics lessons. The lesson was unusually strong mathematically judging by the ratings on the mathematics-specific instruments, where we see many high ratings. Students were asked to make conjectures about the relationship between head circumference and height measurements, and to explain how they were interpreting the data gathered in class and projected on the screen. The mathematics-specific instruments varied in their emphasis on mathematical practices and mathematical accuracy, with MQI having multiple elements regarding accuracy and TRU having none. And when it came to ratings on classroom management, the ratings were low as the lesson lacked crisp transitions, and the teacher struggled to gain student attention at times especially as students were busy gathering measurements in pairs.

Element Density

Some instruments have more specification of a particular construct than others. When observers differ in their ratings for a construct with a small set of elements, those differences can be more pronounced—and create differences in the low-medium-high ratings. With more elements in a construct, differences in individual ratings are reduced. We attribute this difference to what we call "Element Density."

To show the variation in ratings across different instruments, we consider Construct 4: *Lesson Design, Coherence, and Implementation*, and how two instruments pick up on different facets of this lesson. We located the "mathematical focus, coherence, and accuracy" dimensions of the TRU instrument in our Construct 4*: Lesson Design, Coherence, and Implementation*, because in the TRU instrument these dimensions are cast in mathematically specific terms. During a whole group discussion, the highest score on the TRU instrument was assigned in line with the instrument specification: "The mathematics discussed is relatively clear and correct, AND connections between procedures, concepts and

contexts (where appropriate) are addressed and explained" (Schoenfeld, 2013). Amidst the buzz of the classroom activity, one student, Lena, raised her hand and said, "I don't get it...how me and Brady are the same [value for the decimal equivalent of their height-to-circumference ratio]. I have, like, 149 [centimeters, for her height] and 52 [centimeters, for her head circumference]; and Brady has 151 and 53." The teacher immediately recorded Lena's thinking on the whiteboard as the question: "How do we have the same decimal equivalents if we are different heights?" After allowing the students to work for a few more minutes, the teacher then addressed the whole class, re-voicing Lena's question.

In contrast to the high degree of content specification in TRU, we compare the Marzano domains that pick up elements of Construct 4. Marzano, an instrument for all subject areas, includes fifteen different domains that are strongly related to Construct 4. While the TRU instrument focuses the rater's attention squarely on the mathematical substance of a lesson and its implementation, in the Marzano instrument the mathematical substance is taken up across multiple domains. For example, using the Marzano instrument a rater would assign a high score to a lesson where a clear learning goal and scale for self-monitoring has been made explicit, where students are organized to interact, practice, and deepen understanding of new knowledge, and where a lively pace and enthusiasm are maintained. These features were not especially visible in this lesson so while it was mathematically very strong using the TRU instrument, on the Marzano instrument it was middling. Specific content concerns—such as "connections between procedures, concepts and contexts" in the TRU instrument—are mostly absent in the Marzano instrument. Broad descriptions of content learning— "deepening understanding of new knowledge"— are found in the Marzano instrument. It is also important to note that high rankings (4 or 5 on a scale of 5) on the Marzano instrument are reserved for instruction where the majority of students are monitored for evidence of meeting each of the domain's goals, or where instruction is adapted based on that monitoring to better help students meet those goals. Note again, there were only two items in TRU associated with this construct as opposed to fifteen

in Marzano. When we calculate overall scores for this construct, those fifteen items produced a lower score using Marzano due to the different emphases in these two instruments. This last example helps illustrate how the element density *within* a particular construct (e.g. lesson design, coherence, and implementation) for one instrument, in this case, Marzano, can cause variation in ratings. We note, by the way, that neither Marzano nor Danielson, the other non-mathematics specific instrument, produced a single "high" rating, specifically within the subject matter constructs. This example illustrates how ratings can vary across instruments.

Returning to our theoretical framework based on the Mathematical Knowledge for Teaching (Ball et al., 2008) and the Model of Effective Teaching Behaviour (Maulana et al., 2017), we see subject-matter features in high relief in mathematics-specific instruments; other features, such as management, recede. Correspondingly, we see the Marzano instrument's heavy attention to lesson design and implementation with 15 elements connected to Construct 4. However, only two out of 60 elements are associated with the quality of the academic task, aligned with Construct 2, and this likely explains differential scoring. The first element in Marzano, Domain 21: Organizing Students for Cognitively Complex Tasks, focuses on students generating and testing hypotheses. For this particular lesson, the teacher organized the students into pairs with the goal of generating and testing a hypothesis (the relationship between head circumference and height), but instead data collection and arithmetic computations occupied most of the students' time. There was little opportunity for students to engage in the high-quality task of hypothesis generation and testing (the second element of Marzano, Domain 22: Engaging Students in Cognitively Complex Tasks). Moreover, the Marzano instrument focuses only on generating and testing hypotheses, whereas instruments such as MQI and M-Scan bring more detail about mathematical tasks. For example, the following element from MQI resulted in a high score on this same lesson:

> Students engage with content at a high level of cognitive activation. Examples of cognitively activating activities include when students:
>
> - Determine the meaning of mathematical concepts, processes, or relationships
> - Draw connections among different representations or concepts
> - Make and test conjectures
> - Look for patterns
> - Examine constraints
> - Explain and justify (Learning Mathematics for Teaching Project, 2011, p. 20).

Students did not need to draw conclusions among different representations since only one representation was discussed publicly during the lesson. However, students did make meaning of the mathematical concept, make and test conjectures, look for patterns in the table, and explain and justify their thinking. The presence of these elements led to the high score using the MQI instrument.

Similarly, we see in the Head/Height Lesson that some mathematics-specific instruments highlighted different standards of mathematical practices, process standards, and/or the strands of mathematical proficiency while general instruments often did not. For instance, regarding the use of representations, Marzano has one element: "The teacher engages students in activities that help them record their understanding of new content in linguistic ways and/or represent the content in nonlinguistic ways" (Marzano, 2013, p. 14). In contrast, M-Scan has three sub-elements in the representation domain:

> Use of Representations: The extent to which the lesson promotes the use of and translation among multiple representations (pictures, graphs, symbols, words) to illustrate ideas and concepts. The use of and translation among representations should allow students to make sense of mathematical ideas or extend what they already understand.

- Presence of Representations: Teacher and/or students often use more than one representation for a mathematical concept.
- Teacher Translation among Representations: For the representation(s) used, the teacher often makes connections to concepts and between representations.
- Student Translation among Representations: Students translate back and forth between representations. They also explain their representations at times (Berry et al., 2010, p. 10).

Even though both instruments yielded similar ratings in Construct 3: Mathematical Practices, where use of representations was recognized, the Marzano instrument specifies differentiation between verbal and nonverbal presentations of content only. M-Scan, on the other hand, focuses on the presence and nature of mathematical representations in order to advance student understanding.

Overall, we found that the type of observational instrument used for evaluation made little difference on the overall scores for most lessons. However, when instruments did produce different overall scores, it mattered; and we classified these differences by instrumental occlusion, instrumental emphasis, and element density.

Discussion

We found that different instruments rarely yielded different evaluations of the same lesson overall, and when they did, the range was fairly narrow: nearly always low versus medium, or medium versus high, with no discernible pattern across instruments. Of the ten coded lessons, only one, the Head/Height Lesson, was scored as low versus high on seven of the nine constructs. For this lesson, the mathematics-specific instruments produced higher ratings than the instruments intended for all subject areas. This discrepancy across instruments deserves attention even if occurring in only 10% of lessons, because that difference can have serious consequences for a teacher's

evaluation, cast doubts about the face validity of evaluation systems, and undermine instructional improvement efforts.

Not surprisingly, the mathematics-specific instruments provide expectations for mathematics instruction that are more detailed than the general instruments. The mathematics-specific instruments differ from one another as well, in terms of the dimensions of teaching and learning mathematics which they emphasize. These differences highlight discrepancies in perspective around pedagogy, content, and the nature of the evaluation of teaching (Hill et al., 2007).

Our findings suggest that the choice of observation instrument does not usually render much difference in the assessment of most lessons. Yet when the instrument does make a difference, it can differ quite a bit. Again, the Head/Height Lesson detailed in this article received high ratings for the quality of task on mathematics-specific evaluation instruments but received low scores on the same construct when evaluated using general instruments. This should give pause to those using teacher observation for high-stakes employment decisions.

Instrument choice is crucial if a school system intends to target particular areas for improvement. It follows that the improvement of general pedagogical expertise—say, asking high-quality instructional questions—might be best served by an observation instrument that could be used across all subjects. Similarly, if the aim is to improve content-specific dimensions of instruction, such as the quality of mathematical tasks, then a mathematics-specific instrument would be a better choice. Here we highlight the use of these instruments towards instructional improvement. The variation across instruments can guide instrument selection towards instructional improvement.

This study's limitations include the small number of lessons that were rated, as well as the selection of instruments that were used. The instruments used in this study have since been revised; we hypothesize that similar findings would be obtained even with revised instruments. New instruments have likely been developed and will continue to be. Other instruments can expand our understanding and prove useful for researchers and practitioners alike; rating a large number of lessons across multiple instruments is needed in further research. Furthermore,

in our study, not all observers were trained in all instruments, and only two members of the research team had conducted teacher observations in school settings. The constructs developed in this study to compare ratings across instruments sometimes required the inclusion of an element in a construct that did not fully represent every aspect of the element. It also bears mentioning that rating lessons involves interpretation, and scores can vary across raters or even by the same rater scoring at different times. Finally, we did not calculate inter-rater reliability estimates and instead preserved all original ratings for all lessons.

More research is needed comparing, in a much more detailed way, the different observation instruments that are used across the country, as well as how raters are trained and the relationship of instrument to the background knowledge and experience of observers. Understanding the interplay between curriculum materials and observed instruction, with special attention to the agency that teachers may or may not have over curricular choices, is especially needed and could be explored in future research. The quality of an observed lesson may be in part a function of mandated and scripted curriculum materials in use, sometimes under strict pacing regimes, and our observation instruments do not account for this. In this study, raters did not know how teachers came to use particular tasks or sequences of activity.

Our research indicates that overall ratings of observed lessons are rarely affected by the choice of instrument. But in some cases, instrument choice might produce divergent ratings and feedback, especially when it comes to the teaching of specific content. This study suggests that observation instruments might be productively employed for establishing shared language for apprehending and improving instruction, and instrument choice can be directed towards targeted areas of need. Teacher education both for preservice and in-service teachers could make use of such instruments as frameworks for specific feedback about instruction. Shared observation instruments across educator preparation institutions and the school districts they supply could lend cohesion to the teacher

learning experience and build expertise over time. Additionally, analysis and comparison of multiple content-specific instruments would give preservice and in-service teachers a chance to reflect on the nuanced differences that theoretical perspectives shape in instruction. Returning to our theoretical framework, the analysis of general *and* content-specific instruments could offer opportunities for educators to widen their conceptualizations of quality instruction. Who observes whom, whether observation data are positioned as descriptive material for shared investigation or evidence amassed for judgment, the degree to which teachers direct their own learning over time—these are among the many issues that must be attended to with great care in using observation instruments for teacher learning and growth.

In practice, a single observation is not sufficient for teacher learning or evaluation. This study suggests that multiple observations and a full complement of data sources for appraising teachers is warranted. Van der Lans, van de Grift, van Veen, and Fokkens-Bruinsma (2016), for example, found that "reliable feedback requires at least three lesson visits by three different observers and that reliable summative decisions require more than 10 visits" (p. 88). Discerning use of these various instruments, in ways that are directed towards teachers' needs and the specific aims of a school or district, has the potential to improve instruction and provide useful data as part of a mosaic of components in teacher evaluation. Typically, systems rely on a single instrument and a limited number of observations even when there can be significant variation in these ratings returned by different instruments. Although it remains a serious challenge to train raters to use even a single instrument reliably (Lewis et al., 2020), one novel (albeit unwieldy) approach might be to use multiple instruments to compensate for the systematic and random gaps in individual instruments. Our research leads us to advise schools to make intentional choices of observation instruments in alignment with explicit purposes for their use, elevating raters' and teachers' voices as part of that process.

References

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389-407. https://doi.org/10.1177/0022487108324554

Berry, III, R. Q., Rimm-Kaufman, S. E., Ottmar, E. M., Walkowiak, T. A., & Merritt, E. (2010). *The Mathematics Scan (M-Scan): A measure of mathematics instructional quality*. Unpublished measure, University of Virginia. Retrieved from http://www.socialdevelopmentlab.org/wp-content/uploads/2012/05/M-Scan_measure_Final.pdf

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, *22*(2), 71-94. https://doi.org/10.1080/10627197.2017.1309274

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593-2632. https://doi.org/10.1257/aer.104.9.2593

Danielson, C. (2013). *The Framework for Teaching evaluation instrument.* The Danielson Group.

Darling-Hammond, L., Wei, R. C., & Johnson, C. M. (2009). Teacher preparation and teacher learning: A changing policy landscape. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of Education Policy Research* (pp. 613-636). Routledge. https://doi.org/10.4324/9780203880968.ch48

Desimone, L., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1–22. https://doi.org/10.3102/01623737026001001

Eco, U., & Weaver, W. (1994). *How to travel with a salmon & other essays*. Harcourt Brace.

Erickson, F. (2006). Definition and analysis of data from videotape: some research procedures and their rationales. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 177-191). Routledge. https://doi.org/10.4324/9780203874769-15

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record, 107*(1), 186-213.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition & Instruction, 26*(4), 430-511. https://doi.org/10.1080/07370000802177235

Hill, H.C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371-384. https://doi.org/10.17763/haer.83.2.d11511403715u376

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts? In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 111–155). Information Age.

Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance*. Center for Public Education. Retrieved from http://www. centerforpubliceducation.org.

Kersten, T. A., & Israel, M. S. (2005). Teacher evaluation: Principals' insights and suggestions for improvement. *Planning and Changing, 36*(1/2), 47-67.

Learning Mathematics for Teaching Project (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education, 14*(7), 25-47.

Lewis, J. M., Reid, D., Bell, C., Jones, N. D., & Qi, Y. (2020). The mantle of agency: Principals' use of teacher evaluation policy. *Leadership and Policy in Schools*. https://doi.org/10.1080/15700763.2020.1770802

Lord, B. (1994). Teachers' professional development: Critical colleagueship and the role of professional communities. In N. Cobb (ed.), *The future of education: Perspectives on national standards in America* (pp. 175-204). College Entrance Examination Board.

Marzano, R. J., Toth, M. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement.* ProQuest Ebook Central.

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2017). Validating a model of effective teaching behaviour of pre-service teachers. *Teachers and Teaching*, *23*(4), 471-493. https://doi.org/10.1080/13540602.2016.1211102

Millman, J. (1981). *Handbook of teacher evaluation*. Beverly Hills, CA. Sage Publications.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. National Council of Teachers of Mathematics.

No Child Left Behind Act of 2001, Pub. L. 107-110, 20 U.S.C. § 6319 (2002).

Popham, W. J. (2013). On serving two masters: Formative and summative teacher evaluation. *Principal Leadership*, *13*(7), 18-22.

Sawada, D. & Piburn, M. (2000). *Reformed teaching observation protocol (RTOP)*. Arizona Collaborative for Excellence in the Preparation of Teachers.

Schoenfeld, A. (2013). Classroom observations in theory and practice. *ZDM Mathematics Education, 45*(4), 607-621. https://doi.org/10.1007/s11858-012-0483-1

Van der Lans, R. M., van de Grift, W. J. C. M., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluative decisions based on classroom observation. *Studies in Educational Evaluation, 50*, 88-95. https://doi:10.1016/j.stueduc.2016.08.001