Urban Education
2017, Vol. 52(7) 829–861

DOI: 10.1177/0042085915574528
journals.sagepub.com/home/uex
SAGE

Middle School Mathematics Instruction in Instructionally Focused Urban Districts

Melissa D. Boston[1] and Anne Garrison Wilhelm[2]

Abstract

Direct assessments of instructional practice (e.g., classroom observations) are necessary to identify and eliminate opportunity gaps in students' learning of mathematics. This study examined 114 middle school mathematics classrooms in four instructionally focused urban districts. Results from the Instructional Quality Assessment identified high percentages of lessons featuring cognitively challenging tasks, but declines in cognitive challenge during implementation and discussions. Overall instructional quality exceeded results from studies with nationally representative samples and paralleled results of studies of instructionally focused urban middle schools. Significant differences existed between districts, favoring the district with veteran teachers, long-term use of *Standards*-based curricula, and professional development initiatives.

Keywords

mathematics, middle school, programs, school effectiveness, urban education

In more than a decade of educational policy advocating standardized testing as the primary means of improving mathematics teaching and learning, students in U.S. classrooms continue to post substandard performance on mathematical

[1]Duquesne University, Pittsburgh, PA, USA
[2]Southern Methodist University, Dallas, TX, USA

Corresponding Author:
Melissa D. Boston, School of Education, Department of Instruction and Leadership in Education, Duquesne University, 600 Forbes Ave., Pittsburgh, PA 15282, USA.
Email: bostonm@duq.edu

assessments at state, national, and international levels (National Research Council, 2011). As evidence, only 30% of U.S. eighth-grade students scored high or advanced in mathematics on the 2011 Trends in International Mathematics and Science Study (TIMSS), compared with a minimum of 61% of students at high or advanced levels in the five top-performing countries (Mullis, Martin, Foy, & Arora, 2012). In the 2011 National Assessment of Educational Progress (NAEP), only 35% of the national sample of eighth graders demonstrated mathematical proficiency (National Center for Education Statistics [NCES], 2011a), with only Massachusetts having more than half (51%) of eighth-grade students proficient in mathematics. Substandard performance on mathematical achievement tests is even more pronounced in urban and rural schools, schools serving large populations of students with limited English proficiency, and schools in areas with high poverty (U.S. Department of Education, NCES, 2006). In the NAEP 2011 Trial Urban District Assessment, which analyzed data from a subset of 21 urban districts participating in the 2011 NAEP, an average of 26% of eighth-grade students demonstrated mathematical proficiency (NCES, 2011b).

Across this same time period, results from educational research consistently indicate that the most significant factors associated with students' mathematical achievement are pedagogical (Boaler & Staples, 2008; Hiebert et al., 2003; Stein & Lane, 1996). Differences in the implementation of curricula and other resources, between schools and between teachers within the same school, provide different opportunities for learning that subsequently generate differences in student achievement. Even when high-quality resources are present, student learning is mainly affected by how resources or curricula are *implemented* in the classroom. Understanding how to improve students' opportunities to learn mathematics thus requires direct assessments, based on observations and artifacts of teaching, of what teachers and students are doing in classrooms in the process of teaching and learning mathematics. This is particularly important in urban districts, where a deep understanding of students' opportunities to learn mathematics is essential for identifying strengths of the system (i.e., classroom practices that appear to be supporting students' learning) and pathways for improvement (i.e., classroom practices that might be changed to enhance students' learning).

Toward this purpose, mathematics education research consistently identifies a set of instructional practices that appear to support students' learning of mathematics with understanding, collectively called "ambitious mathematics instruction" (Franke, Kazemi, & Battey, 2007). Research connecting ambitious mathematics instruction to student achievement has identified key instructional components, such as cognitively challenging tasks (i.e., tasks that engage students in making sense of mathematics; Hiebert et al., 2003;

Stein & Lane, 1996; Tarr, Reys, Reys, Chavez, Shih, & Osterlind, 2008) and mathematical discussions (Boaler & Staples, 2008), and has delineated specific ways teachers enact or implement these practices successfully (e.g., McClain, 2002; Stein, Engle, Smith, & Hughes, 2008). Nationally commissioned reports (e.g., Kilpatrick, Swafford, & Findell, 2001) and standards from the National Council of Teachers of Mathematics (NCTM; 2000) and the Common Core State Standards in Mathematics (National Governors Association, 2010) advocate ambitious mathematics instruction. Mathematics curricula designed to support such instruction (see Kilpatrick, 2003) are now widely available by commercial publishers. Hence, current research, standards, and curricula can equip districts to implement ambitious instruction, and students' opportunities to learn mathematics can be assessed by identifying a set of well-defined instructional practices through direct observations of teaching.

In this investigation, we utilize classroom observations to examine middle school mathematics instruction in four large urban school districts. The districts were participating in the *Middle School Mathematics and the Institutional Setting of Teaching* (MIST) project,[1] which investigated how differences in school and district settings influence mathematics teachers' instructional practices and students' mathematics achievement over a 4-year period (Cobb & Smith, 2008). Each district was committed to significant educational reforms in middle school mathematics, driven by the goal of enhancing students' learning and understanding of mathematics. Because of their intention to increase students' scores on standardized tests by improving classroom instruction, rather than (and often antithetical to) concentrated efforts to "teach to the test" (Le, Lockwood, Stecher, Hamilton, & Martinez, 2009), we refer to these districts as "instructionally focused." They faced challenges typical of large urban districts (e.g., large percentages of families in poverty, high rates of student and teacher turnover), but were atypical in their approach to improving mathematics teaching and learning.

We explore the following research questions using data from the first year (2007-2008) of the project:

Research Question 1: What is the rigor of instructional tasks, task implementation, and mathematical discussions in urban middle school classrooms?

Research Question 2: What opportunities do students in urban middle school classrooms have to engage in mathematical discussions?

Research Question 3: Are there differences between districts in the study in students' opportunities to learn mathematics in urban middle school classrooms?

Research Question 4: How do the results of this study compare with the results of previous studies that also used classroom observations to assess ambitious mathematics instruction in urban schools?

In the next section, we describe how direct assessments of instruction are necessary for understanding students' opportunities to learn mathematics, and we summarize previous studies that utilized classroom observations to identify ambitious mathematics instruction.

Background

The NCTM (2000) Equity Principle states, ". . . all students need access each year they are in school to a coherent, challenging mathematics curriculum that is taught by competent and well-supported mathematics teachers." However, differences in instructional quality between school districts in the United States with different demographic and socioeconomic conditions are well documented, as "the opportunity gap in students' access to qualified teachers between students of high and low socioeconomic status (SES) was among the largest in the world" (Akiba, LeTendre, & Scribner, 2007, p. 369): Students in low-SES categories (e.g., qualifying for free/reduced lunch) and ethnic minority groups (e.g., Black, Hispanic) are more likely than their high-SES, non-minority peers to (a) have novice teachers; (b) have uncertified or out-of-field teachers, particularly in mathematics; and (c) attend schools with high teacher instability.

Instructional quality can also vary greatly within a school, as different teachers create dramatically different learning environments for students. Disparities in students' opportunities to learn are intensified by highly qualified and experienced teachers often selecting or being assigned to teach advanced mathematics classes, resulting in unequal rates of academic growth for students depending on which teachers and level of mathematics classes they are assigned (McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Rowan, Correnti, & Miller, 2002). Combining low-quality teaching in remedial mathematics classes with disproportionate numbers of children from minority, poor, or English-learning subgroups assigned to such classes, the cycle of substandard performance is reinforced and perpetuated. Underserved populations remain underserved, and children who need the most mathematical support and the best mathematics instruction do not receive it.

The prevalence of and reliance on product-oriented accountability (i.e., student achievement scores and/or gains in scores over time) have limited the development and use of a process-oriented system focused on students' opportunities to learn and capable of characterizing the teaching and learning

that occurs in classrooms, schools, or districts. Comparisons of students' achievement disaggregated across race, socioeconomic status, English proficiency, or other demographic factors do not provide parallel comparisons of students' *opportunities to learn* within the classroom or school settings (Lipman, 2004). In other words, what children are able to achieve is not benchmarked against what they have the opportunity to achieve. By elevating the importance of students' opportunities to learn, differences among subgroups and schools may be easier to explain and eradicate by shifting the focus to "the *conditions* of learning as well as the outcomes" (Gutstein et al., 2005, p. 93; emphasis in original). Milner (2010) suggested, "(a)s an explanation of disparate outcomes, *opportunity* is multifaceted, complicated, process-oriented, and much more nuanced than achievement" (p. 7, emphasis added). Hence, direct assessments of students' opportunities to learn mathematics, through observations or artifacts of teaching, capable of capturing the activities in which teachers and students engage during mathematics instruction, are needed to deeply examine and improve students' *outcomes* in learning mathematics (Pianta & Hamre, 2009; Stein & Matsumura, 2008).

In this study, we utilized classroom observations to identify students' opportunities to learn mathematics in large urban districts. Consistent with Perry (2013), we consider *opportunity to learn* specific to mathematics teaching and learning, defined by (a) the nature of mathematics instructional tasks and (b) how tasks are implemented during instruction, including opportunities for mathematical discussions. We conceptualize *opportunity gaps*, specific to students' learning of mathematics, as differences in opportunities to learn mathematics created or perpetuated by the choice of mathematics instructional tasks and nature of task implementation and discussion (i.e., differences due to the presence or absence of ambitious mathematics instruction). In the next section, we describe components of ambitious mathematics instruction and justify why this framework provides important indicators of students' opportunities to learn mathematics.

Ambitious Mathematics Instruction

The conceptualization of students' opportunities to learn mathematics by considering instructional tasks, task implementation, and discussion is informed by research originating with the Quantitative Understanding: Amplifying Student Achievement and Reasoning (QUASAR) project (Silver & Stein, 1996). Stein, Smith, Henningsen, and Silver (2009) defined a *mathematical task* as a mathematical problem or set of problems that address a related mathematical idea or context, and they distinguish between cognitively challenging ("high-level") tasks and rote, procedural ("low-level")

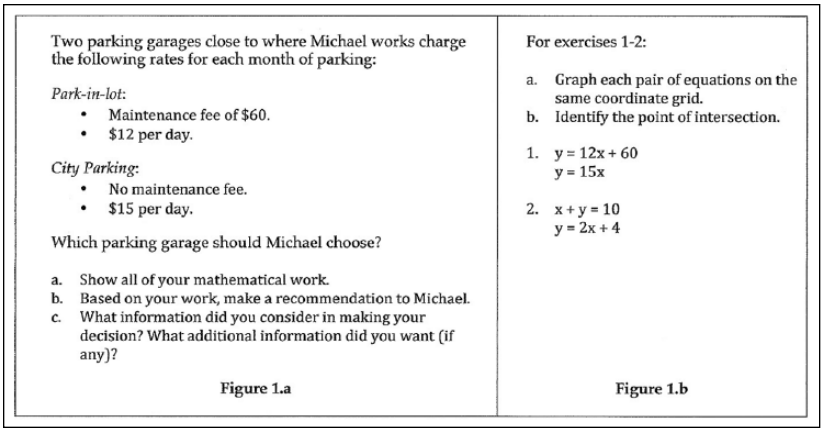

Figure 1. Tasks with different levels of cognitive challenge.

tasks. *Cognitively challenging tasks* provide students opportunities to engage in problem solving, thinking and reasoning, and/or developing an understanding of mathematical ideas, procedures, and formulas (Stein, Grover, & Henningsen, 1996). Rote or procedural tasks engage students in reproducing or practicing facts, procedures, or computations without connection to meaning or understanding. Figure 1 provides an example of a high-level and low-level task to engage students in finding the solution to a system of linear equations, represented by the point of intersection on the graph of the equations. In Figure 1a, the task engages students in problem solving, mathematical modeling (of the parking garage costs), and decision making. The task does not suggest a solution strategy, and students could solve the task using tables, graphs, equations, or reasoning about the context. The task in Figure 1b provides students only procedural practice in graphing linear equations and identifying the point of intersection. We are not suggesting that students do not need to memorize mathematical facts or practice mathematical procedures. We assert, however, that students need greater opportunities to explore and understand mathematics by engaging in cognitively challenging mathematical work and thinking, and that this type of work provides greater access, interest, and opportunity to learn.

Task implementation refers to ways in which tasks are enacted by teachers and students during mathematics lessons (i.e., how teachers support students' work on mathematical tasks and how students actually engage with the mathematics). In the Mathematical Task Framework, Stein and colleagues (1996) described how task challenge can change from (a) the task as it appears in print, (b) the task as set up or introduced by the teacher, and (c) the task as

implemented by the teacher and students during the lesson. In *ambitious mathematics instruction*, teachers introduce (or "launch") a cognitively challenging task and maintain the challenge during implementation by (a) supporting students to engage with (or "explore") the task and (b) orchestrating whole-group discussions where students share mathematical work and thinking, justify claims, make connections between mathematical ideas, and "summarize" the mathematical goals of the lesson (McClain, 2002; NCTM, 2000; Stein et al., 2008).

Ambitious mathematics curricula containing cognitively challenging tasks, such as *Connected Mathematics Project 2* (CMP2; Lappan, Fey, Fitzgerald, Friel, & Philips, 2006) middle school curriculum used in three of four districts in this study, have been shown to increase student performance on problem-solving assessments and minimize achievement gaps, while maintaining students' performance on basic skills and computational assessments (Post et al., 2008; Reys, Reys, Lapan, & Holliday, 2003; Ridgeway, Zawojewski, Hoover, & Lambdin, 2003; Riordan & Noyce, 2001; Schoenfeld, 2002; Thompson & Senk, 2001). Among teachers using ambitious curricula, student achievement is highest in classrooms where students experience consistent opportunities to engage in high-level thinking and reasoning during mathematics instruction. Higher performing students in the United States (e.g., Boaler & Staples, 2008; Stein & Lane, 1996; Tarr et al., 2008) and internationally (e.g., Hiebert et al., 2003) have teachers who sustain students' engagement in cognitively challenging work. Schoenfeld (2002) identified significantly higher achievement among students having teachers rated as "high-implementers" of ambitious elementary and middle school mathematics curricula than students having teachers rated as "low-implementers" in a large urban school district with ambitious goals for mathematics instruction. Specific aspects of ambitious instruction (e.g., setting high expectations, valuing students' efforts, maintaining cognitive challenge, and fostering mathematical inquiry and discussion) appear to affect student achievement and minimize achievement gaps regardless of the type of curriculum in place (Boaler & Staples, 2008; Tarr et al., 2008).

Hence, studies over the past decade relating mathematics teachers' instructional practices to student achievement invariably determine that *teaching matters*. Although ambitious teaching has been associated with improved test scores, often the nature and depth of students' learning cannot be captured on current standardized achievement tests. Scholars have identified the shortcomings of standardized tests as measures of students' mathematical learning (Kilpatrick, 2003; National Mathematics Advisory Panel, 2008) and as measures of teaching quality (Le et al., 2009; McCaffrey et al., 2003). Standardized tests designed to assess a greater depth of mathematical understanding (i.e.,

tests developed for the Common Core State Standards initiative), containing test items beyond memorization and procedures, will be more likely to capture students' learning in ambitious instructional settings. Even so, test scores neither provide data about the aspects of instruction that supported or inhibited students' learning and subsequent test performance, nor can they identify pathways for instructional improvement or disparities in students' opportunities to learn mathematics. Efforts to minimize achievement gaps should grow from efforts to minimize opportunity gaps (Flores, 2007), and identifying differences in students' opportunities to engage in ambitious mathematics instruction is a promising step toward this goal.

We acknowledge that several factors beyond ambitious teaching affect students' opportunities to learn mathematics in urban schools, including (a) the percentage of non-certified mathematics teachers in schools serving African American and low-income students (Jackson & Wilson, 2012); (b) teachers' perceptions of students' mathematical abilities, such as deficit, color-blind, or meritocratic mind-sets (Jackson & Wilson, 2012; Martin, 2007; Milner, 2010); or (c) students' mathematical identities, and how they see themselves (or have been positioned to see themselves) as learners and doers of mathematics (Boaler & Staples, 2008). Ambitious instructional practices can, however, provide a specific framework from which teachers can begin to hold students to higher expectations, provide mathematical work that is engaging and relevant, and develop students' identities as capable mathematicians.

Classroom Observation Studies Identifying Ambitious Mathematics Instruction

Several studies have utilized observations of teaching to assess ambitious mathematics instruction in U.S. classrooms. The TIMSS 1999 Video Study (Hiebert et al., 2003) and the Inside the Classroom Study (Weiss, Pasley, Smith, Banilower, & Heck, 2003) examined nationally representative samples of school districts. Both studies identified a dearth of opportunities for U.S. students to engage in cognitively challenging work in mathematics classrooms. TIMSS observed 100 eighth-grade U.S. mathematics classrooms. Although 15% of instructional tasks could provide opportunities for conceptual understanding, *less than 1%* of tasks were implemented in ways that supported students' development of mathematical concepts (Hiebert et al., 2003). Similarly, Inside the Classroom Study rated only 15% of 364 observed lessons (in K-12 mathematics and science) as high quality based on the criteria of intellectual rigor, teacher questioning for conceptual understanding, and students' opportunities for sense-making (Weiss et al., 2003).

Recently, the Measures of Effective Teaching (MET) Project (Kane & Staiger, 2012) conducted a large-scale study exploring the use of classroom observations, student surveys, and student achievement data to produce a robust measure of teaching effectiveness. In 2009-2010, researchers analyzed 1,000 mathematics lessons in Grades 4 to 8 from public schools across the country. According to the research report (Kane & Staiger, 2012), "scores are highest for competencies related to creating an orderly environment and lowest for those associated with the most complex aspects of instruction" (p. 8). Observed lessons frequently demonstrated content alignment and mathematical accuracy, but infrequently demonstrated ambitious instructional practices, such as student participation in reasoning or investigation, problem-based approaches, and teachers' questioning strategies.

Other studies specifically examined districts utilizing *Standards*-based middle school mathematics curricula (e.g., *Connected Mathematics Project* [CMP]) and/or engaging teachers in professional development. The QUASAR Project (Silver & Stein, 1996) provided professional development to middle school mathematics teachers from five urban districts with economically disadvantaged student populations. Many of these teachers were utilizing pilot versions of current *Standards*-based curricula. Based on a representative sample of 144 observations from 1990 through 1993, with teachers observed for three 3-day cycles yearly, (a) 74% of observed lessons featured cognitively challenging tasks, (b) 31% of observed lessons provided evidence of students engaging in cognitively challenging mathematical work and thinking throughout the lesson, and (c) 50% of observations included discussions where students provided mathematical explanations and justifications (Stein et al., 1996).

The Middle School Mathematics Study observed 33 middle school mathematics teachers in 10 districts, with 2 districts classified as urban by the research team (e.g., serving a city with a population greater than 100,000 people). Researchers compared instructional practices and student achievement between teachers using *Standards*-based curricula (and receiving professional development specifically around using the curricula) and teachers using traditional curricula (and not receiving curriculum-specific professional development). Two observations per teacher indicated that 70% of teachers using *Standards*-based curricula maintained learning environments in which (a) lessons promoted conceptual understanding, (b) lessons supported the exploration of multiple perspectives and strategies, (c) students made mathematical conjectures, (d) students explained their responses or strategies, and (e) teachers used and built upon students' contributions (Tarr et al., 2008). The study does not indicate whether results were consistent across districts or whether differences existed in the urban districts. Consistent

with QUASAR results, teachers with access to *Standards*-based curricula and professional development enacted far greater percentages of high-quality lessons than teachers in national and large-scale samples.

In 2005-2006, the Instructional Leadership Study (Quint, Akey, Rappaport, & Willner, 2007) conducted observations in 49 elementary schools in three urban districts with content-focused professional development ranging from 1 to 5 years. Results from observations of 132 third-grade mathematics lessons identified 65.1% with low overall quality and 14.4% with moderate to high quality when considering instructional tasks, task implementation, and discussion. Results are not disaggregated by teachers' years of professional development support or by the use of *Standards*-based curricula. Overall, these results appear more consistent with national samples than with QUASAR or the Middle School Mathematics Study.

Table 1 provides a summary of characteristics of the highlighted classroom observation studies and the current investigation, listed chronologically according to year(s) of classroom observations. We draw on the results of these studies to situate our findings regarding students' opportunities to learn mathematics, as evidenced by the presence of ambitious mathematics instruction, in urban districts. Next, we describe the methodology in this investigation.

Method

Data for this investigation come from the initial year of a 4-year study investigating what it takes to improve middle-grades mathematics teaching at the scale of four large urban districts. Each year (2007-2011), the MIST project collected several types of data to test and refine hypotheses and conjectures about district and school organizational arrangements, social relations, and material resources that might support mathematics teachers' development of ambitious instructional practices at scale (Cobb & Smith, 2008). This report describes the nature of mathematics instruction during the first year of the project (2007-2008): specifically, (a) the rigor of instructional tasks and task implementation, (b) students' opportunities to engage in mathematical discussions, (c) differences in students' opportunities to learn mathematics between districts, and (d) comparisons to previous classroom observation studies that assessed ambitious mathematics instruction.

Sample

Table 2 provides student demographic information for the four study districts. All four districts serve a significant number of non-White students, and more than half of the students in each district receive free or reduced-price

Table 1. Summary of Characteristics of Classroom Observation Studies.

Study	Classroom observation year(s)	Urban school districts (*n*)	Number of observations	Middle-grades students	*Standards*-based curriculum	Professional development
QUASAR	1990 to 1993	Yes (5)	144	Yes (Grades 6-8)	Yes	Yes
TIMSS	1999	National sample	100	Yes (Grade 8)	NA[a]	NA
Inside the Classroom Study	2000 to 2002	National sample	364	No (Grades K-12)	NA	NA
Middle School Mathematics Study	2003 to 2004	Partially (2 of 10)	66	Yes (Grades 6-8)	Yes	Yes
Instructional Leadership Study	2005 to 2006	Yes (3)	132	No (Grade 3)	NA	Yes
MIST Year 1	2008	Yes (4)	114	Yes (Grades 6-8)	Some (3 of 4)	Some (1 of 4)
MET Project	2009 to 2010	Yes (6)	1,000	Yes (Grades 4-8)	NA	NA

Note. QUASAR = Quantitative Understanding: Amplifying Student Achievement and Reasoning; TIMSS = Trends in International Mathematics and Science Study; MET = Measures of Effective Teaching.
[a]NA indicates that the portion of schools using *Standards*-based curricula or providing professional development cannot be determined from the description of the study.

lunches. Although typical of large urban districts in the challenges they face, including serving large numbers of traditionally low-performing students in mathematics, having limited resources, under-prepared teachers, and high teacher turnover (Darling-Hammond, 2000), these districts are atypical in their instructionally focused approach to increasing student achievement in middle school mathematics. All four districts share the vision of ambitious mathematics instruction and intend to improve student achievement in middle school mathematics by supporting teachers' development of ambitious instructional practices (as opposed to teaching to the test).

The project team purposefully selected a sample of schools from each district with the goal of choosing schools representative of the district as a whole, while selecting approximately 30 teachers per district. Given the variation in school size, the sample of schools ranged from 6 in District C to 10 in District A. Within each school, we created a randomly ordered list of mainstream mathematics teachers. We then offered study participation in that order and recruited two to five middle school mathematics teachers per school. Given the voluntary nature of the study, we had some schools where the first set of randomly selected teachers agreed to participate and other

Table 2. Student Demographic Information for Districts A, B, C, and D.

District	Number of students	% White	% Black	% Hispanic	% LEP	% Free/ reduced price lunch
A	35,000	30	40	15	10	65
B	80,000	15	25	60	30	70
C	160,000	15	30	65	35	85
D	95,000	55	35	5	5	55

Note. LEP = Limited English Proficient; To protect the anonymity of the districts, the number of students is rounded to the nearest 5,000 and percentages are rounded to the nearest 5%.

Table 3. Demographic Information for Participating Teachers.

District	Number of teachers	Mean years of experience teaching math	% White	% Black	% Fully certified
A	28	13.2	89.3	3.6	100
B	26[a]	8.9	69.2	19.2	80.8
C	28	9.2	24.0	62.1	93.1
D	32[a]	8.7	84.4	12.5	87.5

[a]The number does not represent the full sample from the district. We do not have demographic information for three teachers (two in District B and one in District D) who participated in the study.

schools where we had to ask multiple teachers to find the desired number of willing participants. Because of our sampling approach, it is possible that the sample of teachers is not perfectly representative of the teaching population in each district, though it is also likely that our sample does not differ dramatically. Table 3 contains demographic information for participating teachers, by district.

As shown in Table 3, teachers in District A are significantly more experienced than teachers in the other three districts ($p < .05$). Another critical difference between districts (not reflected in Table 3) is the curriculum: District C is the only district in our study that had not adopted CMP2 as its primary curriculum. Instead, the primary curriculum was procedural in nature, and teachers were expected to supplement with CMP2. Furthermore, Districts B and D were in their first year of implementation of CMP2, whereas District A had a long-standing commitment to high-quality curriculum (including CMP2) and professional development initiatives.

Measuring Ambitious Mathematics Instruction

Data on teachers' instructional practices and students' opportunities to learn mathematics were collected using the Instructional Quality Assessment (IQA) Lesson Observation rubrics (Boston, 2012) for Academic Rigor (AR) and Accountable Talk (AT). The constructs measured by the IQA rubrics align with the ambitious curricular and instructional reform efforts undertaken by districts in this study. Boston (2012) provided a thorough conceptual foundation for the IQA rubrics, described briefly here.

Academic Rigor (AR). Stein and colleagues' (1996) Mathematical Tasks Framework and Levels of Cognitive Demand served as the main conceptual framework for the AR rubrics:

- *Task Potential* assesses the level of cognitive demand of the main instructional task (i.e., the task that occupied the most instructional time in the lesson). This dimension is rated by considering the level of thinking required to produce a complete and thorough response that satisfies the stated demands of the task.
- *Implementation* assesses teacher's implementation of and students' engagement with the instructional task. While *Task Potential* assesses the level of rigorous thinking that the task has the *potential* to elicit from students, *Implementation* assesses the level of rigorous thinking in which students *actually* engaged. The score for this dimension is holistic, reflecting the highest level of engagement of the majority of students during individual or small-group work on the task and during any discussion following students' work on the task.
- *(Rigor of the) Discussion* assesses the level of cognitive processes evident in the discussion following students' work on the task (i.e., whether students show their work and/or explain their thinking about important mathematical content). This dimension provides an overall, holistic rating of the discussion on how the talk advances students' understanding of the *mathematical* content. The discussion contributes to the *Implementation* score, and also receives a separate score for *Discussion*.

Each dimension of AR is rated on a scale of 0 to 4 (0 indicates the construct was absent) that represents a continuum of low to high levels of rigor and reflects discrete categories of cognitive demand. As summarized in Figure 2, the descriptors for each score level are relatively consistent across dimensions, though the referent changes from mathematical tasks (*Task*

Score Level		Academic Rigor Rubrics *Potential, Implementation, Discussion*	Accountable Talk Rubrics *Teacher Linking, Student Linking, Teacher Press, Student Providing*
High-Level	4	High-level cognitive demands (Stein et al. 2009) with an explicit explanation, generalization, proof, or connection required by the task (*Potential*), exhibited by students (*Implementation*), or evident in the discussion (*Discussion*).	At least 3 occurrences of the talk move talk move is used consistently throughout the whole-group discussion.
	3	High-level cognitive demands (Stein et al. 2009), but the explanation, generalization, proof, or connection is not explicitly required by the task or is implicit in students' work, explanations, or discussion.	At least 2 strong occurrences of the talk move; talk move is present but not used consistently throughout the whole-group discussion.
Low- Level	2	Tasks, students' work, explanations, or discussions are procedural, algorithmic, or computational.	Weak or formulaic occurrences of the talk move; or one strong occurrence of the talk move.
	1	Tasks, work, or discussions that involve memorization, recall, taking notes, or single-word responses.	The discussion lacked the specific talk move.
	0	No task, implementation, or discussion was present in the observed lesson.	No discussion was present in the observed lesson.

Figure 2. Summary of IQA score levels.

Potential), to the cognitive processes evident throughout the lesson (*Implementation*), to the cognitive processes evident in the discussion (*Discussion*). This rating scheme facilitates comparisons across dimensions and fosters a strong *qualitative* idea of what each score level "looks like" in an actual classroom situation.

Accountable Talk (AT). AT (Resnick & Hall, 2001) upholds the standards of the discipline of mathematics for accuracy, evidence for claims, and reasoning behind ideas and conjectures, *while also* responding to, developing, and advancing the knowledge, ideas, and claims of all students in the classroom (i.e., talk that is accountable to the discipline and the learning community).

In the IQA rubrics, AT is measured through *linking* and *press*[2]:

- "Linking" describes talk that is accountable to the learning community: revoicing (O'Connor & Michaels, 1996); prompting students to extend, analyze, or critique the mathematical work and thinking of others (Cazden, 2001; Cobb, Boufi, McClain, & Whitenack, 1997; McClain, 2002); and students' connections and comparisons to the work or ideas of others. Raters consider whether the teacher makes explicit talk moves to support students in connecting ideas and positions to build coherence in the discussion (*Teacher's Linking*), and whether student's contributions explicitly link to and build on each other (*Students' Linking*).
- "Press" describes talk moves accountable to the discipline: teachers' prompting students to explain their thinking, validate the accuracy of their computations, and justify their claims (Boaler & Staples, 2008). *Teacher's Press* and *Students' Providing* assess teachers' press for accurate knowledge, thorough explanations of ideas, and appropriate justification for claims in classroom talk, and students' efforts to provide accurate knowledge and evidence to support their claims, present arguments, and draw conclusions.

The AT rubrics are rated solely on the whole-group discussion following students' work on the task(s), and not on any talk that occurs during the introduction of the task or as students work (individually or in small groups) on the task itself. The frequency and quality of talk moves for each score level are held consistent across the *Linking* and *Press* rubrics. A score of 0 indicates that no discussion occurred. Score levels 1 and 2 reflect low-quality AT: the absence of a talk move or type of student response, weak or minimal attempts to make the talk move by teachers, or weak or minimal attempts to link ideas together or justify their knowledge and claims by students. Score levels 3 and 4 represent consistent, high-quality AT moves and student responses. The consistency of high-quality scores (3 or 4) versus low-quality scores (1 or 2) within the AT rubrics and between the AT and AR rubrics enhances the interpretive value of the IQA rubric results.

Data Collection and Analysis

Throughout January to March 2008, videographers recorded 2 days of instruction (consecutively, when possible, to account for lessons that might extend beyond 1 day) for each participating teacher. Teachers were asked to include a problem-solving activity and a related whole-group discussion in

the observed lessons. To be clear, the goal of the video-recordings was not to capture the nature of teachers' *everyday* practice, but rather to assess the extent to which a teacher might enact the particular kind of instruction articulated by district leaders as the goal of the instructional reform. Given the directions to include a problem-solving lesson and whole-group discussion, it is consistent to consider what was video-recorded as teachers' best shot at enacting ambitious instructional practices.

The video-recorded lessons were coded using the IQA Lesson Observations rubrics for AR and AT. The IQA rubrics were tested for reliability and validity by the project team (Matsumura, Garnier, Slater, & Boston, 2008)) and external researchers (Quint et al., 2007). Coders were trained to use the IQA rubrics reliably. Before actual coding began, coders achieved 80% reliability on previously coded videos, chosen to represent the variety of anomalies that coders would encounter. Each coder was then randomly assigned a list of teachers. The set of two class-days for each teacher was coded chronologically, given that lessons from the first day might continue into the second day (resulting in one set of scores for the spanning lesson). Over the course of the coding period, one set of teacher scores for each coder was randomly checked for reliability once every 2 weeks to account for rater drift, which resulted in double-coding of approximately 15% of the lessons. When differences in scores occurred on the double-coded lessons, coders reached a consensus through discussion. The overall percent exact-point agreement in initial coding was 71.3% with an average kappa score[3] of 0.49. Consensus scores were used in all analyses.

We analyzed one set of IQA rubric scores per teacher. For teachers with two complete sets of rubric scores (e.g., teachers who completed two entire lessons over the 2 days of videotaping), we consistently selected for analysis the highest set of scores over the 2 days of instruction. Recall that we perceive the video-recorded lessons as teachers' best shot at enacting ambitious instruction, because we did not record teachers frequently enough to capture typical classroom practice. Our decision for choosing the highest set of IQA rubric scores for each teacher is consistent with this perception. For example, in cases when teachers only had whole-class discussions on one of the two days of instruction, we selected the set of scores for the lesson involving a whole-class discussion. When teachers' lessons spanned both days of videotaping, we assumed this 2-day lesson was the teacher's best effort (especially given the extended time to enact the lesson).

Consistent with our intent to characterize the nature of instruction across four large urban school districts striving for ambitious mathematics instruction, we present descriptive statistics for each district, analyze differences in students' opportunities to learn mathematics between districts, and compare

our results with previous research. By providing information about distributions and standard deviations, we provide an indication of within-district variation. With our emphasis on district trends, however, we focus the analyses on district means and distributions, comparisons between districts, and district-level comparisons to previous research. With score levels 0 to 4 representing a scale of increasing quality *and also* distinct categories of performance, mean scores are provided to support interpretations of rubric results within a district, and non-parametric tests provide comparisons between districts. We describe particular tests used with corresponding results in the following section.

Results

Given our interest in the nature of middle-grades mathematics instruction at the scale of four large, urban school districts, we examined scores on the rubrics for approximately 30 teachers in each district for the 2007-2008 school year. Table 4 provides district means, standard deviations, and score frequencies for each rubric. We draw on data in Table 4 to characterize and compare students' opportunities to learn mathematics across the four districts.

AR: Instructional Tasks, Task Implementation, and Discussion

First, we highlight findings pertaining to Research Question 1 and the mathematical rigor of instructional tasks, task implementation, and whole-group discussions.

Tasks. In Districts A, B, and D, *Task Potential* means were 3.14, 3.17, and 3.18, respectively. *Task Potential* means above 3 indicate the use of cognitively challenging tasks during the majority of observed lessons, also evident in the percent of lessons scoring a 3 or 4 for *Task Potential* in Districts A, B, and D (85.7%, 82.1%, 72.8%).[4] More than half of observed lessons in District C (62.1%) featured high-level instructional tasks, though District C posted the lowest task mean (2.66) and the lowest percent of lessons (3.5%) with instructional tasks scoring a 4 for *Task Potential* (i.e., cognitively challenging tasks that explicitly required students to provide, explain, or illustrate mathematical thinking and reasoning), with Districts A, B, and D at 28.6%, 35.7%, and 45.5%, respectively.

Implementation. *Implementation* means in each district fell below a score of 3 (2.67, 2.32, 2.03, 2.58). District A exhibited high-level instruction in the majority of observed lessons, with 53.6% of lessons scoring 3 or 4 in

Table 4. Quality of Observed Mathematics Instruction by District: Academic Rigor and Accountable Talk Rubrics, 2007-2008 School Year.

		Number (%) at each score level				
	M (*SD*)	0	1	2	3	4
District A (*n* = 28 teachers)						
Task potential	3.14 (0.65)	0 (0)	0 (0)	4 (14.3)	16 (57.1)	8 (28.6)
Implementation	2.67 (0.72)	0 (0)	0 (0)	13 (46.4)	11 (39.3)	4 (14.3)
Discussion	2.21 (1.03)	2 (7.1)	4 (14.3)	10 (35.7)	10 (35.7)	2 (7.1)
Teacher linking	2.04 (0.79)	2 (7.1)	1 (3.6)	20 (71.4)	4 (14.3)	1 (3.6)
Student linking	1.25 (0.80)	2 (7.1)	20 (71.4)	4 (14.3)	1 (3.6)	1 (3.6)
Teacher press	2.32 (1.02)	2 (7.1)	2 (7.1)	12 (42.9)	9 (32.1)	3 (10.7)
Student providing	1.93 (1.05)	2 (7.1)	8 (28.6)	10 (35.7)	6 (21.4)	2 (7.1)
District B (*n* = 28 teachers)						
Task potential	3.17 (0.72)	0 (0)	0 (0)	5 (17.9)	13 (46.4)	10 (35.7)
Implementation	2.32 (0.48)	0 (0)	0 (0)	19 (67.9)	9 (32.1)	0 (0)
Discussion	1.75 (0.80)	2 (7.1)	7 (25)	15 (53.6)	4 (14.3)	0 (0)
Teacher linking	1.79 (0.69)	2 (7.1)	4 (14.3)	20 (71.4)	2 (7.1)	0 (0)
Student linking	1.04 (0.43)	2 (7.1)	23 (82.1)	3 (10.7)	0 (0)	0 (0)
Teacher press	1.93 (0.90)	2 (7.1)	5 (17.9)	15 (53.6)	5 (17.9)	1 (3.6)
Student providing	1.79 (0.88)	2 (7.1)	6 (21.4)	18 (64.3)	0 (0)	2 (7.1)

(continued)

Table 4. (continued)

		Number (%) at each score level				
	M (*SD*)	0	1	2	3	4
District C (*n* = 29 teachers)						
Task potential	2.66 (0.55)	0 (0)	0 (0)	11 (37.9)	17 (58.6)	1 (3.5)
Implementation	2.03 (0.42)	0 (0)	1 (3.5)	27 (93.1)	0 (0)	1 (3.5)
Discussion	1.27 (1.00)	7 (24.1)	10 (34.5)	10 (35.7)	1 (3.5)	1 (3.5)
Teacher linking	1.48 (0.91)	7 (24.1)	2 (6.9)	19 (65.5)	1 (3.5)	0 (0)
Student linking	.79 (0.49)	7 (24.1)	21 (72.4)	1 (3.5)	0 (0)	0 (0)
Teacher press	1.55 (1.09)	7 (24.1)	5 (17.3)	11 (37.9)	6 (20.7)	0 (0)
Student providing	1.52 (0.99)	7 (24.1)	3 (10.3)	16 (55.2)	3 (10.3)	0 (0)
District D (*n* = 33 teachers)						
Task potential	3.18 (0.85)	0 (0)	0 (0)	9 (27.3)	9 (27.3)	15 (45.5)
Implementation	2.58 (0.71)	0 (0)	0 (0)	18 (54.6)	11 (33.3)	4 (12.1)
Discussion	1.82 (1.01)	5 (15.2)	5 (15.2)	14 (42.4)	9 (27.3)	0 (0)
Teacher press	1.67 (0.92)	5 (15.2)	5 (15.2)	20 (60.6)	2 (6.06)	1 (3)
Student linking	.97 (0.53)	5 (15.2)	24 (72.7)	4 (12.1)	0 (0)	0 (0)
Teacher asking	1.73 (1.18)	5 (15.2)	9 (27.3)	13 (39.4)	2 (6.1)	4 (12.1)
Student providing	1.58 (0.97)	5 (15.2)	10 (30.3)	12 (36.4)	6 (18.2)	0 (0)

Implementation. Districts B and D exhibited high-level *Implementation* in 32.1% and 45.4% (respectively) of observed lessons, whereas high-level *Implementation* occurred in only 3.5% of observed lessons in District C. These data indicate that, with the exception of District A, more than half of observed lessons did not engage students in high-level thinking and reasoning. Instead, as suggested by the percentage of observed lessons scoring a 2 in *Implementation* for all districts in the study, instruction typically focused on procedures without connections to meaning and understanding. Very few lessons received *Implementation* scores of 4: four lessons (14.3%) in District A, one (3.5%) in District C, and four (12.1%) in District D.

While *Task Potential* provides information about the *potential* rigor of the mathematical activity in the classroom, *Implementation* characterizes the *actual* rigor of mathematical activity during the lesson. A Wilcoxon Signed-Rank test suggests that for all four districts, the mean for *Implementation* is significantly lower than the mean for *Task Potential*, indicating a decline in cognitive challenge between students' *opportunities for* and *actual engagement in* thinking and reasoning during the observed lessons overall ($z = 7.67$; $p < .001$) and within each district ($z = 3.35$, $z = 4.19$, $z = 3.75$, and $z = 3.96$, for Districts A, B, C, and D, respectively; $p < .001$).

Rigor of the discussion. The majority of observed lessons in all districts exhibit low-quality mathematical discussions. District A was the highest among the districts, with a mean of 2.21 and 42.8% of *Discussion* scores at 3 or 4. Districts B, C, and D posted mean scores below 2, with low percentages of high-level discussions (14.3%, 7.0%, 27.3%). Only three observed lessons scored 4 for *Discussion*, two in District A (7.1%) and one in District C (3.5%). At the low end of the scale, scores of 0 or 1 (indicating no discussion or discussion characterized by one-word responses) occurred in more than half (58.6%) of lessons in District C, but less than one third of lessons in the other districts (21.4%, 32.1%, 30.4%).

AT: Students' Opportunities to Engage in Mathematical Discourse

Given the general low quality of whole-group discussions, it is not surprising that findings pertaining to Research Question 2 and opportunities for rich mathematical discussions are equally sparse. Across all districts, few instances occurred of the AT constructs *Linking* and *Press*. This suggests that even when teachers conduct whole-group discussions, students are rarely given opportunities to connect to each other's mathematical work and thinking or to offer rich mathematical explanations and justifications.

Teacher and student linking. Minimal occurrences of *Linking* occurred during the observed lessons. Means for *Teacher Linking* were at or below 2 in each district (2.04, 1.79, 1.48, 1.67) and means for *Student Linking* near or below 1 (1.25, 1.04, 0.79, 0.97). District A again posts the highest scores, with high-level (scores of 3 or 4) *Teacher Linking* and *Student Linking* occurring in 17.9% and 7.2% of observed lessons, respectively. High-level *Teacher Linking* was demonstrated infrequently in Districts B, C, and D (7.1%, 3.5%, 9.0%). No lessons in Districts B, C, or D exhibited high-level *Student Linking*. Lessons receiving a 4 were limited to two occurrences of *Teacher Linking* (one [3.6%] in District A and one [3.0%] in District D) and one occurrence of *Student Linking* (District A, 3.6%).

Teacher press and student providing. Instances of teachers pressing for students' reasoning and justification and of students providing valid reasons and justifications also occurred infrequently during the observed lessons. Only District A posted a mean score for *Teacher Press* above 2 (2.32, 1.93, 1.55, 1.73), whereas all districts' mean scores for *Student Providing* fell below 2 (1.93, 1.79, 1.52, 1.58). In District A, high-quality *Teacher Press* occurred in 42.8% of observed lessons, approximately twice as often as other districts (21.5%, 20.7%, 18.2%). High-quality *Student Providing* occurred in 28.5% of observed lessons in District A, followed by 18.2% in District D, 10.3% in District C, and 7.1% in District B. Score level 4 was achieved for *Teacher Press* in eight lessons overall: three (10.7%) in District A, one (3.6%) in District B, and four (12.1%) in District D. Four lessons reached a 4 in *Student Providing*: two each in Districts A (7.1%) and B (7.2%).

Differences Between Districts

Table 4 provides district means and standard deviations. Statistically significant differences between districts ($p < .05$; see Figure 3), identified using Wilcoxon Rank-Sum tests (also called the Mann-Whitney two-sample statistic), occurred when (a) District A outscored District C on all rubrics except *Student Providing* and (b) District C fell significantly lower than Districts B and D on *Task Potential, Discussion*, and *Implementation*. No significant differences existed between Districts A and B or between B and D, and Districts A and D differed significantly only on *Teacher Press.*

Across rubrics, no significant differences occurred between districts on *Student Providing*, and no significant differences were found between Districts B, C, and D on all AT rubrics. Significant differences occurred mainly on AR rubrics, with District C significantly lower than other districts on all AR rubrics.

	A v. B	A v. C	A v. D	B v. C	B v. D	C v. D
Task Potential		A>C (z=2.81)		B>C (z=2.81)		D>C (z=2.62)
Implementation		A>C (z=4.06)		B>C (z=2.82)		D>C (z=3.69)
Discussion		A>C (z=3.33)		B>C (z=2.12)		D>C (z=2.28)
Teacher Linking		A>C (z=2.37)				
Student Linking		A>C (z=2.48)				
Teacher Press		A>C (z=2.53)	A>D (z=2.31)			
Student Providing						

Figure 3. Comparisons of district mean scores in 2007-2008.
Note. Empty cells denote no statistically significant difference between districts. Cell contents give direction of significant difference with *z* scores in parenthesis ($p < .05$).

Comparisons With Previous Classroom Observation Studies

Table 1 enables comparisons between the results of this study (of four large, instructionally focused urban districts), and previous results from nationally representative or large-scale samples of districts (e.g., TIMSS, Inside the Classroom Study, MET Project) and from instructionally focused urban districts (e.g., QUASAR, Middle School Mathematics Study, Instructional Leadership Study). First, lessons in this study generally exhibited higher levels of instructional quality than nationally representative samples of districts in the TIMSS and Inside the Classroom studies, where no more than 15% of observed lessons demonstrated high-quality tasks and/or implementation. For districts in this study, percentages of lesson observations featuring cognitively challenging instructional tasks (rated 3 or 4) ranged from 62.1% to 85.8%. Only District C had fewer than 15% of *Implementations* considered high-level, and Districts B and C had fewer than 15% of *Discussions* rated highly. While the MET study did not provide exact percentages of ambitious instructional practices, researchers "rarely found highly accomplished practice . . . associated with the intent to teach students higher-order thinking skills" (Kane & Staiger, 2012, p. 10). In the current study, though we rarely identified high-quality discussions, notable percentages (32.1%-53.6%) of lesson implementation in three districts engaged students in higher order thinking skills.

Several aspects of observed instruction in this study were consistent with results from studies of instructionally focused urban middle schools with professional development and/or *Standards*-based mathematics curricula (e.g., QUASAR, Middle School Mathematics Study). The Middle School Mathematics Study identified 70% of lessons with overall high quality. QUASAR researchers identified 74% of lessons with high-level tasks, 31% of lessons with high-level implementation, and 50% of lessons with high-level discussions. All districts in our study posted comparable percentages (62.1%-85.7%) of high-level tasks in the observed lessons. Districts A, B, and D exceeded QUASAR in percentage of high-level implementations (53.6%, 32.1%, 45.4%). Percentages of high-level discussions in District A (42.8%) approached the percentage observed in QUASAR (50%), with other districts falling below 25%. District C, dissimilar to other districts in this study in the lack of a *Standards*-based curriculum, differed considerably from QUASAR and the Middle School Mathematics Study in high-level implementation (3.5%) and discussions (7%).

The Instructional Leadership Study (Quint et al., 2007) utilized the same rubrics as this investigation. Means on the AR rubrics across all observations were 2.26 for *Task Potential*, 2.10 for *Implementation*, and 1.76 for *Discussion*. In this investigation, observed lessons demonstrate the same pattern in mean scores as the Instructional Leadership Study, with the highest mean occurring for *Task Potential* and consistently lower means in *Implementation* and *Discussion*. Districts A, B, and D posted higher means than districts in the Instructional Leadership Study on all AR rubrics (except *Discussion* in District B). Both studies identified similarly low means on the AT rubrics.

Although all districts in the current investigation were aiming for ambitious mathematics instruction, the observed lessons indicate significant variation in teacher's enactment of such instruction. In the discussion that follows, we describe what the results indicate about students' opportunities to learn mathematics in urban middle school classrooms.

Discussion

Ambitious mathematics instruction provides opportunities for students to learn mathematics with understanding, and has been shown to decrease achievement gaps (Boaler & Staples, 2008; Schoenfeld, 2002). Hence, identifying components of ambitious mathematics instruction can provide a means for identifying differences, or opportunity gaps, in how mathematics is taught and learned in different districts, schools, and classrooms. Only by attending to these opportunity gaps can we begin to eradicate achievement gaps (Flores, 2007; Lipman, 2004).

In this discussion, we use our results to characterize middle school mathematics instruction in four instructionally focused urban districts. Consistent with the larger goals of the project, we hypothesize how differences in instructional practices and students' opportunities to learn mathematics connect to aspects of the institutional setting (namely, availability of *Standards*-based mathematics curricula and professional development opportunities), and we suggest pathways for improvement. We also situate our findings within prior research to assess the progress of ambitious mathematics instruction over time.

Characterizing Mathematics Instruction in Instructionally Focused Urban Districts

Potential of the task. The majority of observed lessons featured cognitively challenging instructional tasks, and with the exception of District C (3.5%), notable percentages of tasks in Districts A (28.6%), B (35.7%), and D (45.5%) explicitly required students to provide, explain, or illustrate their mathematical thinking and reasoning (i.e., *Task Potential* score of 4). By posing instructional tasks with high cognitive demands in the majority of observed lessons, teachers in each district provided students opportunities for mathematical learning and understanding. In the QUASAR study, the consistent presence of high-level instructional tasks, regardless of the level of implementation, resulted in moderate gains in student achievement (Stein & Lane, 1996).

Across several studies, including TIMSS (Hiebert et al., 2003), QUASAR (Stein & Lane, 1996), and studies using the IQA rubrics (Boston, 2012; Boston & Smith, 2009; Quint et al., 2007), tasks with low cognitive demands are rarely implemented in ways that result in high-level thinking and reasoning. In the majority of observed lessons, *Task Potential* sets the ceiling for *Implementation*, and in fact, for all discussion-based rubrics as well. Tasks with low cognitive demands simply do not provide fodder for teachers to engage students in thinking, reasoning, or mathematical discourse throughout the enactment of the lesson. If opportunities for high-level thinking and reasoning are not embedded in instructional tasks, these opportunities rarely materialize during mathematics lessons. This finding, robust in its consistency across several studies, suggests that *Standards*-based curricula and/or high-level instructional tasks are a necessary condition for ambitious mathematics instruction. Cognitively challenging tasks can support positive mathematical identities by positioning students as learners and doers of mathematics, setting high expectations, providing multiple access points, and encouraging multiple solution strategies—features of the instructional setting noted as particularly important for the success of African American and

low-income students in urban schools (Boaler & Staples, 2008). As part of a framework for considering students' opportunity to learn mathematics, *using cognitively challenging instructional tasks* can provide a concrete first step for teachers to elicit and recognize students' mathematical abilities and perhaps begin to move beyond deficit, innate-ability, or meritocratic mind-sets.

Task implementation. In all districts, *Implementation* means were significantly lower than *Task Potential* means, indicating that students' actual engagement in thinking and reasoning during the observed lessons did not reflect the opportunities for high-level cognitive processes embedded in instructional tasks. This decline suggests that students are not fully benefitting from opportunities for mathematical learning in *Standards*-based curricula or high-level tasks. Empirical research from more than a decade indicates that the highest learning gains occur in classrooms where students consistently engage in high-level thinking and reasoning (e.g., Schoenfeld, 2002; Stein & Lane, 1996; Tarr et al., 2008). Hence, opportunity gaps and achievement gaps could be affected to a greater extent if high-level cognitive demands were consistently maintained during implementation. High-level *Implementation* involves holding students accountable for the mathematical work and thinking in the task and providing students the right amount of support to maintain students' engagement (without taking over the mathematical work and thinking; Henningsen & Stein, 1997). Hence, teachers' instructional moves to maintain high-level demands during implementation can promote positive mathematical identities, establish trust, and communicate high expectations. These classroom practices are identified as particularly important for fostering the success of African American and low-income students in urban schools (Boaler & Staples, 2008; Milner, 2010).

Significant differences in *Implementation* also existed between districts. At the extremes, more than half of observed lessons in District A (15/28; 53.6%) received high-level *Implementation* scores compared with only one lesson in District C (1/29; 3.5%), providing students in each district with distinctly different opportunities to learn mathematics. What aspects of the institutional setting may have affected differences in implementation between districts? Research has identified many challenges in maintaining high-level demands in mathematics classrooms where students and teachers are accustomed to rote procedures and memorization (rather than exploration, thinking, and reasoning; Henningsen & Stein, 1997). More experienced teachers, long-term use of CMP, and professional development initiatives may have affected implementation in District A. Studies of *Standards*-based curricula identify improvements in teachers' implementation and in student achievement over time, with significant increases typically occurring in and beyond

the second year of use (Bray, 2005; Post et al., 2008; Reys et al., 2003), and generally associate "a longer implementation in the school . . . with a greater score advantage for students" (Riordan & Noyce, 2001, p. 383). These findings suggest the importance of persevering with *Standards*-based curricula and maintaining an instructional focus at the administrative level, as districts frequently discard or replace initiatives that do not yield immediate results.

Research also indicates the necessity and value of professional development initiatives in implementing *Standards*-based mathematics curricula (Senk & Thompson, 2003). Teachers may have neither experienced ambitious instruction as learners of mathematics, nor explored or practiced this type of pedagogy during preservice teacher training or field-based experiences (Franke et al., 2007). Professional development specifically aimed at enhancing teachers' ability to enact ambitious instruction and maintain the demands of cognitively challenging instructional tasks has proven effective and enduring (Boston & Smith, 2009; 2011), and is readily accessible in professional development materials (e.g., Smith, Silver, & Stein, 2005; Stein et al., 2009).

Mathematical discussions. The majority of discussions in all districts consisted of students demonstrating procedures or providing brief responses to teachers' questions (i.e., *Discussion* scores of 1 or 2), with few occurrences of the AT constructs of Linking and Press and extremes occurring again between Districts A and C. Notably, almost half (12/28; 42.8%) of discussions in District A were characterized by explanations of students thinking and reasoning (i.e., *Discussion* scores of 3 or 4) and high-level *Teacher Press.* Similar elements of the school setting hypothesized for differences in districts' performance on the *Implementation* rubric can be posited for differences on the rubrics assessing classroom discourse. Teachers in District A had more experience and training in implementing CMP2 and the components of ambitious instruction it entails, including orchestrating whole-group discussions. Hence, similar arguments can also be waged for the value of ongoing professional development opportunities, even within District A, to support teachers to engage students in mathematical discourse.

Comparisons With Prior Research

As shown in Table 1, studies utilizing national samples exhibit a low occurrence of ambitious mathematics instruction, even with the passing of a decade between TIMSS in 1999 and MET in 2009-2010. In contrast, when comparing districts with ambitious mathematics curricula and professional development initiatives over a similar span of time, QUASAR (1990-1993) and the Middle School Mathematics Study (2003-2004) identified far greater percentages of

lessons exhibiting ambitious instructional practices. This finding was not replicated in the Instructional Leadership Study, though the longevity of professional development in each district was unclear. In our study, similar distinctions in ambitious mathematics instruction appear between District A, Districts B and D, and District C. Similar results across District A, QUASAR, and the Middle School Mathematics Study suggest that *Standards*-based curriculum and professional development opportunities are necessary conditions for enacting ambitious mathematics instruction.

Given our sample of four large urban districts with ambitious goals for mathematics instruction, atypical of many urban districts in their instructional focus, it is reasonable to assume that the instructional patterns in our results are the same or more rigorous than what might be found in other large urban districts across the United States. This suggests that the majority of students in urban districts have few opportunities to engage in high-level thinking and reasoning in mathematics, and indicates the need for additional work in providing richer opportunities to learn mathematics for students in urban schools.

Conclusion: Implications for Minimizing Opportunity Gaps

In this investigation, we assessed middle school mathematics instruction in four large urban districts participating in a long-term project seeking to identify how school and district settings affect mathematics teachers' instructional practices and students' learning. We proposed, consistent with recent work by Perry (2013), that cognitively challenging tasks and high-level task implementation provide a useful framework for considering students' opportunities to learn mathematics.

Several hypotheses follow from our work. First, *Standards*-based curricula and/or cognitively challenging instructional tasks appear to be necessary conditions for supporting higher levels of AR. Task levels set the ceiling for the level of implementation *and* for all discussion-based rubrics. Hence, districts and classrooms lacking high-level instructional tasks in mathematics offer students far different opportunities to learn mathematics than classrooms and districts utilizing such tasks. Second, teachers need support to (a) maintain students' opportunities for thinking, reasoning, and problem solving throughout lesson implementation, and (b) orchestrate high-quality whole-class discussions that include AT moves. Results from District A and results of other studies of instructionally focused urban middle schools suggest that professional development experiences may equip teachers to achieve high levels of implementation and discussion. Third, although transience of reform initiatives and teachers frequently plague urban districts, longevity of use of

Standards-based curricula and teachers' experience with such curricula appear to support the successful enactment of ambitious mathematics instruction. Hence, urban districts need to remain instructionally focused, even in the absence of immediate gains in achievement test scores.

More broadly, our work indicates how direct assessments of instructional quality, based on observations and artifacts of teaching, might equip urban districts to (a) monitor reform efforts, including curricular implementation or professional development; (b) identify differences in students' opportunities to learn mathematics; and (c) suggest pathways for providing rich mathematical learning experiences capable of reducing opportunity gaps and achievement gaps.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The empirical work reported in this paper has been supported by the National Science Foundation under grants DRL-0830029 and ESI-0554535. Anne Wilhelm's contributions to the article were supported by the Institute of Education Sciences (IES) pre-doctoral research training program, grant number R305B080025. The opinions expressed do not necessarily reflect the views of either Foundation or IES. We would like to thank Paul Cobb, Thomas Smith, and Glenn Colby for their contributions to this work.

Notes

1. The Middle School Mathematics and the Institutional Setting of Teaching (MIST) study described herein was supported by the National Science Foundation (Paul Cobb and Thomas Smith, Co-PIs; Award No. ESI 0554535). The opinions expressed are those of the authors and do not represent the views of or the National Science Foundation.
2. The IQA also contains an Accountable Talk rubric for Participation not discussed in this article.
3. Kappa (Cohen, 1960) is an adjusted percent agreement measure, based on the proportion of codes in each category. There are no standards for evaluating kappa scores but Hartmann and colleagues (2004) suggest that kappa scores between 0.6 and 0.75 are good.
4. Multiple scores in parentheses represent districts in alphabetical order.

References

Akiba, M., LeTendre, G. K., & Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational Researcher, 36*, 369-387.

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Record, 110*, 8-9.

Boston, M. D. (2012). Assessing the quality of mathematics instruction. *Elementary School Journal*, *113*, 76-104.

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.

Bray, M. S. (2005). *Achievement of eighth grade students in mathematics after completing three years of the Connected Mathematics Project* (Doctoral dissertation). The University of Tennessee, Knoxville. Retrieved from http://trace.tennessee.edu/cgi/viewcontent.cgi?article=3292&context=utk_graddiss

Cazden, C. B. (2001). *Classroom discourse: The language of teaching and learning* (2nd ed.). Portsmouth, NH: Heinemann.

Cobb, P., Boufi, A., McClain, K., & Whitenack, J. (1997). Reflective discourse and collective reflection. *Journal for Research in Mathematics Education, 28*, 258-277.

Cobb, P., & Smith, T. M. (2008). District development as a means of improving mathematics teaching and learning at scale. In K. Krainer & T. Wood (Eds.), *Participants in mathematics teacher education: Individuals, teams, communities, and networks* (Vol. 3, pp. 231-254). Rotterdam, The Netherlands: Sense Publishers.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Darling-Hammond, L. (2000). New standards and old inequities: School reform and the education of African American students. *Journal of Negro Education, 69*, 263-287.

Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap? *The High School Journal, 91*, 29-42.

Franke, M. L., Kazemi, E., & Battey, D. (2007). Mathematics teaching and classroom practice. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 225-256). Greenwich, CT: Information Age.

Gutstein, E., Middleton, J. A., Fey, J. T., Larson, M., Heid, M. K., Dougherty, B., . . . Tunis, H. (2005). Equity in school mathematics education: How can research contribute? *Journal for Research in Mathematics Education, 36*, 92-100.

Hartmann, D. P., Barrios, B. A., & Wood, D. D. (2004). Principles of behavioral observation. In M. Hersen (Series Ed.), *Comprehensive handbook of psychological assessment: Vol. 3. Behavioral assessment* (pp. 108–137). Hoboken, NJ: John Wiley & Sons.

Henningsen, M., & 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.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K., Hollingsworth, H., Jacobs, J., . . . Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 Video Study* (NCES Rep. No. 2003-013). Washington, DC: National Center for Education Statistics.

Jackson, K., & Wilson, J. (2012). Supporting African-American students' learning of mathematics: A problem of practice. *Urban Education, 47*, 354-398.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Research paper: MET Project). Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kilpatrick, J. (2003). What works. In S. L. Senk & D. R. Thompson (Eds.), *Standards-based mathematics curricula: What are they? What do students learn?* (pp. 471-488). Mahwah, NJ: Lawrence Erlbaum.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Philips, E. D. (2006). *Connected mathematics 2*. Boston, MA: Pearson.

Le, V., Lockwood, J. R., Stecher, B. M., Hamilton, L. S., & Martinez, J. F. (2009). A longitudinal investigation of the relationship between teachers' self-reports of reform-oriented instruction and mathematics and science achievement. *Educational Evaluation and Policy Analysis, 31*, 200-220.

Lipman, P. (2004). *Regionalization of urban education: The political economy and racial politics of Chicago-metro region schools*. Paper presented at the annual meeting of the American Education Research Association, San Diego, CA (April, 2004).

Martin, D. B. (2007). Beyond missionaries or cannibals: Who should teach mathematics to African American children? *The High School Journal, 91*, 6-28.

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, *13*, 267-300.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability* (MG-158-EDU). Santa Monica, CA: RAND.

McClain, K. (2002). Teachers' and students' understanding: The role of tool use in communication. *Journal of the Learning Sciences, 11*, 217-249.

Milner, H. R. (2010). *Start where you are, but don't stay there: Understanding diversity, opportunity gaps, and teaching in today's classrooms*. Boston, MA: Harvard Education Press.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

National Center for Education Statistics. (2011a). *The nation's report card: Mathematics 2011* (NCES 2012–458). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Center for Education Statistics. (2011b). *The nation's report card: Trial urban district assessment mathematics 2011* (NCES 2012–452). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association. (2010). *Common core state standards*. Retrieved from http://www.corestandards.org/Math/

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Research Council. (2011). *Incentives and test-based accountability in public education* (Committee on Incentives and Test-Based Accountability in Public Education, M. Hout & S. W. Elliott, Eds.). Washington, DC: The National Academies Press, Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education.

O'Connor, M. C., & Michaels, S. (1996). Shifting participant frameworks: Orchestrating thinking practices in group discussions. In D. Ghicks (Ed.), *Discourse, learning, and schooling* (pp. 63-103). New York, NY: Cambridge University Press.

Perry, A. D. F. (2013). *Equitable spaces in early career teachers' mathematics classrooms* (Unpublished doctor dissertation). North Carolina State University, Raleigh.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observations can leverage capacity. *Educational Researcher, 38*, 109-119.

Post, T. R., Harwell, M. R., Davis, J. D., Maeda, Y., Cutler, A., Andersen, E., . . . Norman, K. W. (2008). Standards-based mathematics curricula and middle-grades students' performance on standardized achievement tests. *Journal for Research in Mathematics Education, 39*, 184-212.

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 from http://www.mdrc.org/publications/470/execsum.html

Resnick, L. B., & Hall, M. W. (2001). *The principles of learning: Study tools for educators* [CD-ROM, Version 2.0]. Pittsburgh, PA: Learning Research and Development Center, Institute for Learning. Available from www.instituteforlearning.org

Reys, R., Reys, B., Lapan, R., & Holliday, G. (2003). Assessing the impact of standards-based middle grades mathematics curriculum materials on student achievement. *Journal for Research in Mathematics Education, 34*, 74-95.

Ridgeway, J. E., Zawojewski, J. S., Hoover, M. N., & Lambdin, D. V. (2003). Student attainment in the connected mathematics curriculum. In S. L. Senk & D. R. Thompson (Eds.), *Standards-based mathematics curricula: What are they? What do students learn?* (pp. 193-224). Mahwah, NJ: Lawrence Erlbaum.

Riordan, J. E., & Noyce, P. E. (2001). The impact of two standards-based mathematics curricula on student achievement in Massachusetts. *Journal for Research in Mathematics Education, 32*, 368-398.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *Teachers College Record, 104*, 1525-1567.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher, 31*, 13-25.

Senk, S. L., & Thompson, D. R. (Eds.). (2003). *Standards-based mathematics curricula: What are they? What do students learn?* Mahwah, NJ: Lawrence Erlbaum.

Silver, E. A., & Stein, M. K. (1996). The QUASAR project: The "revolution of the possible" in mathematics instruction reform in urban middle schools. *Urban Education*, *30*, 476-521.

Smith, M. S., Silver, E. A., & Stein, M. K. (2005). *Improving instruction in algebra: Using cases to transform mathematics teaching and learning* (Vol. 2). New York, NY: Teacher's College Press.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning, 10*, 313-340.

Stein, M. K., Grover, B., & Henningsen, M. (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., & 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 a 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., Henningsen, M., & Silver, E. A. (2009). *Implementing standards-based mathematics instruction: A casebook for professional development* (2nd ed.). New York, NY: Teachers College Press.

Tarr, J. E., Reys, R. E., Reys, B. J., Chavez, O., Shih, J., & Osterlind, S. (2008). The impact of middles grades mathematics curricula on student achievement and the classroom learning environment. *Journal for Research in Mathematics Education, 39*, 247-280.

Thompson, D. R., & Senk, S. L. (2001). The effects of curriculum on achievement in second-year algebra: The example of the University of Chicago School Mathematics Project. *Journal for Research in Mathematics Education, 32*, 58-84.

U.S. Department of Education, National Center for Education Statistics. (2006). *The condition of education 2006* (NCES 2006-071). Washington, DC: U.S. Government Printing Office.

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom*. Chapel Hill, NC: Horizon Research.

Author Biographies

Melissa D. Boston is an associate professor at Duquesne University, where she teaches mathematics education courses for elementary, middle level, and secondary mathematics preservice teachers. She investigates changes in classroom practices of mathematics teachers participating in professional development.

Anne Garrison Wilhelm is an assistant professor at Southern Methodist University where she primarily teaches masters level courses for in-service K-12 mathematics teachers. Her research is focused on understanding mathematics teachers' learning and practice.