Assessing Academic Rigor in Mathematics Instruction: The Development of the Instructional Quality Assessment Toolkit

CSE Technical Report 672

Melissa Boston and Mikyung Kim Wolf
Learning and Research Development Center,
University of Pittsburgh

February 2006

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Bldg., Box 951522
Los Angeles, Ca 90095-1522
(310) 206-1532

Project 2.3 Indicators of Classroom Practice and Alignment
Lauren Resnick and Brian Junker, Project Directors

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

The findings and opinions expressed in this report are those of the author(s) and do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences (IES), or the U.S. Department of Education.

ASSESSING ACADEMIC RIGOR IN MATHEMATICS INSTRUCTION: THE DEVELOPMENT OF THE INSTRUCTIONAL QUALITY ASSESSMENT TOOLKIT

Melissa Boston and Mikyung Kim Wolf
Learning and Research Development Center,
University of Pittsburgh

Abstract

The development of an assessment tool to measure the quality of instruction is necessary to provide an informative accountability system in education. Such a tool should be capable of characterizing the quality of teaching and learning that occurs in actual classrooms, schools, or districts. The purpose of this paper is to describe the development of the Academic Rigor in Mathematics (AR-Math) rubrics of the Instructional Quality Assessment Toolkit and to share the findings from a small pilot study conducted in the Spring of 2003. The study described in this paper examined the instructional quality of mathematics programs in elementary classrooms in two urban school districts. The study assessed the reliability of the AR-Math rubrics, the ability of the AR-Math rubrics to distinguish important difference between districts, the relationships between rubric dimensions, and the generalizability of the assignment collection. Overall, exact reliability ranged from poor to fair, though 1-point reliability was excellent. Even with the small sample size, the rubrics were capable of detecting difference in students' opportunities to learn mathematics in each district. The paper concludes by suggesting how the AR-Math rubrics might serve as professional development tools for mathematics teachers.

Since the release of the No Child Left Behind Act of 2001, the public school accountability system has relied almost exclusively on students' achievement test scores to ascertain instructional quality. The prevalence of this product-oriented accountability has limited the development and use of a process-oriented system capable of characterizing the quality of teaching and learning that occurs in actual classrooms, schools, or districts. Process-oriented assessments that identify elements of instruction that influence students' opportunities to learn could inform policy-makers at the national, state, and district levels; could indicate areas in need of professional development for teaching staff; and could serve as professional development tools for teachers. Supported by a multitude of research linking pedagogy to students' opportunities to engage in high-level thinking and reasoning, the development of an assessment tool to measure the quality of instruction seems

both feasible and necessary to provide an informative accountability system in education.

Until recently, the body of research conducted for the purpose of measuring the quality of instruction in actual classrooms relied predominantly on data obtained from case studies, surveys, or self-reports. Current studies by Horizon Research (Weiss & Palsey, 2004), the TIMSS 1999 Video Study (Hiebert et al., 2003), and CRESST (Borko, Stecher, Alonzo, Moncure, & McClam, 2003; Matsumura, Garnier, Pascal, & Valdes, 2002; Clare & Aschbacher, 2001) have analyzed instructional quality based on classroom observations and artifacts. Similarly, a research team lead by Lauren Resnick, Brian Junker, and Lindsay Clare Matsumura at the University of Pittsburgh's Learning Research and Development Center has endeavored to design the Instructional Quality Assessment (IQA) Toolkit, a set of rubrics that measure the quality of instruction and learning in school language arts and mathematics programs (Junker et at., 2004). The purpose of this report is to describe the development of the IQA's Academic Rigor in Mathematics (AR:Math) rubrics for lesson observations and collections of students' work, to share the findings from a small pilot study conducted in the Spring of 2003, and to posit conclusions from the pilot and future directions for the IQA toolkit. In general, the discussion presented in this report is intended to answer the following question, "Is the IQA toolkit a reasonable means of assessing the academic rigor of school mathematics programs?" Specifically, the research questions addressed in this paper include:

1. How reliable are the AR-Math rubrics?
2. Can the AR-Math rubrics distinguish the quality of mathematics instructional programs?
3. How independent are the dimensions of the AR-Math rubrics?
4. Does the design of the assignment collection in this study provide a valid indicator of the quality of instruction in the observed classrooms?

The discussion begins by describing the theoretical basis of the AR-Math rubrics.

Theoretical Basis of the Indicators of Academic Rigor

In order to measure the quality of instruction, the construct and its indicators must first be defined. That is, the question of "What is quality instruction that support students' learning?" needs to be answered. Based upon a great body of

cognitive and social psychology research, Resnick and her colleagues have established a set of principles of effective teaching and learning called the Principles of Learning (Resnick & Hall, 1998). The Instructional Quality Assessment (IQA) toolkit has been designed to evaluate instructional quality based upon four Principles of Learning that are evident and observable in classrooms that promote students' learning: *Academic Rigor, Accountable Talk, Clear Expectations, and Self-Management of Learning.* The IQA toolkit consists of approximately 20 rubrics accompanied by rater-training materials (Junker et al., 2004). Direct lesson observation (1 per teacher) and collections of classroom assignments (4 per teacher) are used as the major data source for measuring the quality of instruction, and student and teacher interviews are also conducted to provide supplementary information about the observed lessons. The discussion presented in this report focuses on the Principle of Academic Rigor.

Academic rigor in a thinking curriculum holds that students must be exposed to a rich knowledge core that is organized around the mastery of major concepts. This curriculum should provide students with regular opportunities to pose and solve problems, formulate hypotheses, justify their reasoning, construct explanations, and test their own understanding. Students must have opportunities to engage with academically rich content material and to develop their thinking skills in order to achieve at high levels (Institute for Learning, 2002). In mathematics, this Principle can be translated into students' opportunities to learn worthwhile, important mathematics *with understanding*.

Research and theories on learning mathematics with understanding provide insight into academic rigor in mathematics instruction and learning. Constructivist perspectives suggest that learning with understanding occurs as students build on their prior knowledge and actively engage with mathematical ideas in ways that lead to a re-organization of their previous knowledge structures (Romberg & Carpenter, 1986). Hiebert and Carpenter (1992) claim that learning with understanding results as students represent and structure mathematical ideas, both physically and mentally, in ways that facilitate connections between concepts, facts, and procedures. Lesh, Post, and Behr (1987) view mathematical understanding as the ability to recognize a mathematical idea within a variety of representations, to work with the idea within a specific representation, and to translate the idea between different representations. Social-constructivist theories contend that opportunities to learn mathematics with understanding include occasions for

students to collaboratively negotiate, construct, and communicate mathematical ideas and reasoning (Cobb, Boufi, McClain, & Whitenack, 1997; Voigt, 1994). Based on these theories, the National Council of Teacher of Mathematics (NCTM) has released several standards documents portraying a vision of mathematics teaching and learning that promotes mathematical thinking, reasoning, and understanding (NCTM, 2000, 1991, 1989). In this vision, students are to be active constructors of mathematical knowledge, and teachers are to serve as facilitators of students' learning by providing classroom experiences in which students can engage with rich mathematical tasks, develop connections between mathematical ideas and between different representations of mathematical ideas, and collaboratively construct and communicate their mathematical thinking.

NCTM's vision of quality mathematics teaching and learning constitutes our construct of academic rigor in mathematics. In the sections that follow, we present a research base to support our selection of indicators of this construct.

The Influence of Tasks on Students' Learning

The extent to which opportunities are provided for students to learn mathematics with understanding is a key aspect to measure the academic rigor of mathematics instruction. In particular, the cognitive demand of the instructional tasks can be a core indicator reflecting the academic rigor of instruction. A growing body of research supports that curricular materials specifically developed to contain tasks with high-level cognitive demands (United States Department of Education, 1999) are successful in improving students' performance on state and national tests of mathematical achievement (e.g., Fuson, Carroll, & Druek, 2000; Riordan & Noyce, 2001; Schoen, Fey, Hirsch, & Coxford, 1999), in improving students' understanding of important mathematical concepts (e.g., Ben-Chaim, Fey, Fitzgerald, Benedetto, & Miller, 1998; Huntley, Rasmussen, Villarubi, Sangtong, & Fey, 2000; Thompson & Senk, 2001; Reys, Reys, Lapan, & Holliday 2003), and in improving students abilities to reason, communicate, problem-solve and make mathematical connections (e.g., Ridgeway, Zawojewski, Hoover, & Lambdin, 2003; Schoenfeld, 2002).

On a practical level, instructional tasks influence student learning because working on mathematical tasks constitutes what students *do* during the majority of their time in mathematics class (Hiebert et al., 2003). On a theoretical level, Doyle offers two premises for why "tasks form the basic treatment unit in classrooms" (1983, p. 162). First, a mathematical task draws students' attention toward a

particular mathematical concept and provides certain information surrounding that concept (Doyle, 1983). Students are exposed to (and thus have an opportunity to learn) the concepts embedded in the tasks they complete. Students are not exposed to (and thus have much less of an opportunity to learn) content that is not represented in the tasks they complete. Second, tasks influence student learning by setting parameters for the ways in which information about the mathematical concept can be operated on or processed (Doyle, 1983). Students will become skilled at what they have an opportunity to actually do in mathematics class. If students' academic work consists of practicing procedural computations, they are likely to become facile with computational skills; however, if students spend their time reflecting on why things work the way they do, how ideas are connected to their prior knowledge, or how ideas and procedures compare and contrast, then they are likely to be constructing new relationships and new understandings of mathematics (Hiebert et al., 1997).

Hence, different types of tasks provide different opportunities for students' learning and place different expectations on students' thinking. A task that entails only memorization will provide much different opportunities for learning than a task that requires problem-solving, conjecturing, and reasoning. Mathematical tasks with high-level cognitive demands contain features resonant with the perspectives on learning mathematics with understanding noted earlier. For example, high-level tasks often have multiple entry points and solutions strategies, thereby allowing different students to approach the task in different ways based on their own prior knowledge. High-level tasks also feature multiple representations, opportunities to form connections between mathematical ideas or representations, and opportunities for communication (Stein, Grover, & Henningsen, 1996). Hiebert and colleagues (Hiebert et al., 1997; Hiebert & Wearne, 1993) further specify high-level or "appropriate" mathematical tasks as those that provide opportunities for reflection and communication on *important* mathematics, where the mathematics in the task is intellectually challenging, the task connects with students' prior knowledge, and the task leaves behind valuable mathematical "residue" (1997, p. 18). Putnam, Lampert, and Peterson, (1990) contend that high-level tasks involve problem solving, mathematizing (describing a situation in terms of its quantitative relationships), or building mathematical augments. Tasks described as "worthwhile tasks" by NCTM (2000) or as "procedures with connections or "doing mathematics" by Stein, Grover, and Henningsen (1996) feature high-level cognitive demands.

Tasks that are classified as having low levels of cognitive demand involve either memorization or the application of procedures with no connection to meaning or understanding (Stein, Grover, & Henningsen 1996; Doyle, 1983). Tasks with low levels of cognitive demand are not always inappropriate or "bad" instructional tasks. If the goal of an instructional episode is for students to memorize formulae, reproduce a demonstrated example, or practice a given procedure, then tasks that require low levels of cognitive demand are appropriate. However, if the goal of an instructional episode is for students to think, reason, and engage in problem-solving, then instruction must be based on high-level, worthwhile, appropriate mathematical tasks (Stein & Lane, 1996; NCTM, 2000; Hiebert et al., 1997). Hence, the level of cognitive demands of instructional tasks is an important indicator of academic rigor in mathematics instruction.

Task Implementation

The previous section presented the argument that the level of cognitive demand of instructional tasks is an important indicator of quality mathematics instruction. Research has also shown that the level of cognitive demand of a task can be altered over the course of an instructional episode (Henningsen & Stein, 1997). When attempting to implement tasks with high-level cognitive demands, teachers and students accustomed to traditional, directive styles of teaching and procedural tasks can be uncomfortable with the ambiguity and struggle that often accompany high-level tasks (Smith, 1995; Clarke, 1997). In response to the ambiguity or to uncertainty on how to proceed, students may disengage with the task or press the teacher for step-by-step instructions (Romanagno, 1994; Henningsen & Stein, 1997), thereby reducing the cognitive demands of the task as it is implemented during instruction. This tendency is evident in the TIMSS 1999 Video Study, where less than 1% of the lessons in U.S. classrooms in which the instructional tasks could provide opportunities for students to make meaningful mathematical connections (i.e., tasks with high-level cognitive demands) resulted in students actually making those connections during the lesson (Hiebert et al., 2003). Similarly, only 15% of the lessons analyzed by Horizon Research were rated as effectively supporting students' opportunities for learning mathematics (Weiss & Pasley, 2004).

In investigating the link between the implementation of reform-oriented features of mathematics instruction to variations in students' learning, Stein and Lane (1996) found that the greatest student learning gains occurred in classrooms where students were consistently exposed to high-level tasks *and* the high-level

cognitive demands were sustained throughout the lesson. These results appear consistent with findings from the TIMSS 1999 Video Study (Hiebert et al., 2003), in which higher performing countries were found to implement high-level tasks in ways that maintained the high-level cognitive demands. Hence, the level of task implementation appears to be an essential indicator of academic rigor in mathematics instruction.

Mathematical Discussion

One feature that influences the implementation of a mathematical task is students' opportunity to engage in a mathematical discussion following their work on the task. During this discussion, students can see how others approached the task and can gain insight into solution strategies and reasoning processes that they may not have initially considered; teachers can provide opportunities for students to explain their reasoning, make mathematical generalizations, or make connections between concepts, strategies, or representations. The whole group discussion provides an opportunity for teachers to advance the mathematical understandings of *all* students. As noted by Lampert (2001), "In each interaction in a public discussion, a teacher can use a student's connection with some mathematics to teach the student while also teaching the class as a whole" (p. 174). While the discussion may initially focus on the work that students have produced, students should also have opportunities to analyze, compare, connect, and reflect upon the collective mathematical work of the class for a given lesson or task. Referred to as *reflective discourse*, students' work "subsequently becomes an explicit object of discussion" (Cobb, Boufi, McClain, & Whitenack, 1997, p. 258). Opportunities for students to reflect and communicate about their mathematical work are essential for learning mathematics with understanding (Hiebert et al., 1997), and thus serve as an indicator of the quality of mathematics instruction.

The use of generic talk moves, such as linking, revoicing, and press for mathematical evidence and explanations, are also valuable in providing opportunities for all students to learn mathematics during a whole-group discussion (O'Connor & Michaels, 1996). Such talk moves characterize the Principle of Learning entitled Accountable Talk (AT) (Resnick & Hall, 1998), and the IQA Toolkit contains a set of rubrics designed to assess the presence Accountable Talk and how it serves to support students' learning during a lesson (see Wolf, Crosson, & Resnick, 2004). In mathematical discussions, teachers also need to foster students' inquiry in ways that allow the underlying mathematical concepts to become visible. According to

Chazan & Ball (2001), in order for a discussion to advance students' mathematical understanding, a teacher may need to use a combination of generic talk moves and content-specific talk moves that help shape and direct the discourse toward the mathematical goals of the lesson. In this way, the teacher's talk moves provide opportunities for all students to learn from the discussion and ensure that the discussion leaves behind valuable mathematical residue (Hiebert, et al., 1997). Hence, mathematical discussions as an indicator of academic rigor supplements the AT rubrics by assessing the extent to which the talk moves provided opportunities for students to advance their mathematical understandings.

Teacher's Expectations

Implementing high-level tasks in ways that promote students' learning with understanding is often shaped by teachers' and students' beliefs about how mathematics is best taught and learned (Remillard, 1999; Clarke, 1997; Romanagno, 1994). A teacher's perception of the types of learning opportunities that are possible with a given group of students or from a given mathematical task sets parameters around their expectations for students and the level of products and processes for which students are held accountable. According to Schoenfeld, "…teachers' envisionings of what they expect to take place in the classroom play a major role in shaping what does take place" (1998, p. 17). Henningsen and Stein (1997) and Doyle (1988, 1983) identify accountability for high-level products and processes as a factor that contributes to students' sustained engagement with high-level cognitive processes. Students are unlikely to spontaneously go beyond what is required by a task or by a teacher; rather, students will identify the information and operations that are necessary to accomplish the task and will adjust their work strategies to correspond to their perceptions of the task's or the teacher's requirements (Doyle, 1983). Hence, a teacher's expectations for students' learning can influence what happens during an instructional episode, and throughout students' learning experiences in that teacher's classroom. Teachers' expectations determine what students will be held accountable for, and this accountability frames the work that students engage in during instruction. In this way, a teacher's expectations for students' learning serve as an important indicator of academic rigor of mathematics instruction.

In sum, four indicators of students' opportunities to learn mathematics with understanding have been identified: tasks, task implementation, mathematical discussions, and teachers' expectations. These indicators form the basis of the IQA

toolkit's rubrics to measure academic rigor in mathematics instruction. In the next section, we describe the development of the Academic Rigor in Mathematics rubrics.

Development of the IQA Academic Rigor in Mathematics

The IQA Academic Rigor in Mathematics (AR-Math) rubrics were created to consist of four dimensions critical to assessing students' opportunities to learn mathematics with understanding, corresponding to the four indicators discussed in the previous section: the potential of the task, the implementation of the task, students' discussion (this dimension assesses students' written responses for the assignment rubrics), and teachers' expectations. Based on research stemming from the QUASAR project[1] (i.e., Stein, Grover, & Henningsen, 1996; Henningsen & Stein, 1997), all four of these dimensions are rated based on the general notion of high-level and low-level cognitive demands in mathematical tasks (potential), in the cognitive processes evident in the lesson or in student's work (implementation), in the cognitive processes evident in the discussion or in students' written responses to the assignment, and in the level of cognitive processes in the teachers' expectations.

Each rubric uses a 4-point scale (1 = low and 4 = high) that is consistent and generalizable across dimensions. In other words, the descriptors for each score level are relatively constant across dimensions, though the referent changes from task, to task implementation, to students' discussion, to teachers' expectations. This rating scheme facilitates comparisons across dimensions, and enables the classroom observer or the interpreter of the results to develop a strong *qualitative* idea of what each score level "looks like" in an actual classroom situation. Specifically, the descriptors for score levels 3 and 4 are consistent with characteristics of high-level cognitive demands (Stein, Grover, & Henningsen, 1996). Collapsed together, levels 3 and 4 correspond to the highest category used in the analysis of mathematics lessons by TIMSS 1999 Video Study, "making connections" (Hiebert et al., 2003). In our rating scheme, we differentiate Levels 3 and 4 with respect to the complexity and the explicitness of the mathematical connections or reasoning present in the task, the lesson, the discussion, or in the teacher's expectations. Score levels 1 and 2 reflect low-level cognitive demands; with level 2 resonant of "procedures without connections" and level 1 of "memorization" or "no mathematical activity"

[1] The QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning) Project was a national reform project from 1990-1996 aimed at assisting schools in economically disadvantaged communities to develop middle school mathematics programs that emphasized thinking, reasoning, and problem-solving (Silver & Stein, 1996).

(Stein et al., 1996). Similarly, levels 1 and 2 are analogous to the TIMSS categories of "stating concepts" and "using procedures," respectively (Hiebert, et al., 2003). Hence our overall rating scheme is based on the work of Stein and colleagues and is also consistent with the rating scheme used by TIMSS researchers. Note that, in our rating scheme, an important demarcation line exists between score levels 2 and 3 that separates high- and low- level cognitive demands in each dimension of the AR-Math rubrics. Each dimension of the rubric is described more specifically in the paragraphs that follow and, when appropriate, will be compared to the dimensions used by TIMSS (Hiebert, et al., 2003).

AR1: Potential of the Task. For the AR-Math Lesson Observation and AR-Math Assignment rubrics, the "Potential of the Task" dimension assesses the level of cognitive demand that students could potentially engage in by working on the task. The score levels in this dimension are derived from the levels of cognitive demand proposed by Stein, Grover, and Henningsen (1996). This dimension is rated by considering the requirements of the task as written in curricular materials and as introduced by the teacher (especially in primary grades where directions tend to be more verbal that written). Similarly, TIMSS rated the type of instructional tasks according to the categories identified earlier in this section (making connections, using procedures, or stating concepts), and also assessed the complexity of the task (high, moderate, low) and whether tasks were mathematically related (mathematically related, pure repetition, unrelated, or thematically related). As stated earlier in this section, the TIMSS categories correspond to the IQA task levels as follows: Stating Concepts = 1; Using procedures = 2; Making Connections = 3 or 4 depending on the complexity and explicitness of the mathematical reasoning required by the task.

AR2: Implementation. While the Potential of the Task dimension described in the preceding paragraph assesses the level of rigorous thinking that the task has the *potential* to elicit from students, the "Implementation" dimension assesses the level of rigorous thinking that students *actually* engaged in through their work on the task during the lesson or on the assignment. The score for this dimension is holistic, reflecting the highest level at which most of the students engaged with the task throughout the lesson, as they worked on the task and during any whole or small group discussions. Certain instructional factors serve to maintain or to reduce students' opportunities to engage in high-level cognitive processes as they engage with mathematical tasks, and a Mathematics Lesson Checklist based on the factors

identified by Henningsen and Stein (1997) is also provided to guide the scoring of the Implementation dimension. The TIMSS 1999 Video Study (2003) also analyzed task implementation and, following a methodology similar to QUASAR research (e.g., Stein, Grover, & Henningsen, 1996; Stein and Lane, 1996; Henningsen & Stein, 1997), compared the categorization of the task itself (making connections, using procedures, or stating concepts) to the categorization of the task as implemented during the lesson.

AR3: Rigor in Student Discussion or Responses. The dimension of Rigor in the Discussion Following the Task (hereafter referred to as Student Discussion) assesses the level of cognitive processes evident in the discussion (for the Lesson Observation rubric) or in students' written work on the task (for the Assignment rubric). This dimension analyzes whether students show their work and/or explain their thinking about important mathematical content and, for the Lesson Observation rubrics, supplements the Accountable Talk (AT) rubrics by providing an overall, holistic rating of students' talk during the final discussion of the lesson with respect to students' opportunities to learn important mathematical content. While specific (but content-free) AT moves are recorded and assessed on the AT rubrics, this dimension is centered on how the talk advances students' understanding of the *mathematical* content following their work on the task. For example, this dimension assess whether the discussion provides opportunities for reflection, for students to express their reasoning, for students to make connections between concepts, strategies, or representations, or for students to engage in generalizations or proof of mathematical ideas. In parallel, the dimension of Rigor in Students' Responses on the Assignment rubric looks for evidence of these types of cognitive processes in students' written work. For this dimension, the TIMSS analysis of classroom discourse is quite different from the IQA's focus on ascertaining whether the discussion provided opportunities for students to advance their mathematical understandings. Based on videotaped lessons, TIMSS conducted a detailed assessment of specific elements of discourse, at the level of teacher and student utterances, that cannot be captured during live observation.

AR4: Rigor in Teacher's Expectations. This rubric rates the degree of rigorous thinking that the teacher expects throughout the lesson or in the assignment. The teacher's expectations may be conveyed through verbal or written directions, criteria charts, and/or models of exemplary performance that the teacher might share with students. This dimension assesses the level of cognitive demand in the teacher's

expectations for students' work on the task. TIMSS rated teachers' goals for the lesson based on the categories of Skills, Thinking, Social, Test Prep, or "Can't Tell." These categories map onto the AR-Math Teacher's Expectations score levels in as follows: Can't Tell = 0; Social = 1; Skills/Test Prep = 2; Thinking = 3 or 4 depending on the complexity and explicitness of the mathematical reasoning expected by the teacher.

In the next section, we describe how the AR-Math rubrics were incorporated into a pilot study of the Instructional Quality Assessment Toolkit.

The Study

A small pilot study of the IQA was conducted in Spring 2003, using 16 mathematics lessons and 14 reading comprehension lessons from randomly-sampled primary schools in two urban school districts. The study described in this report focuses only on the Academic Rigor in Mathematics rubrics and their implementation. Specifically, the study aimed to investigate whether the IQA toolkit is a reasonable means of assessing the academic rigor of school mathematics programs. The components of the study will be described below.

Participants

In the spring of 2003, a pilot study was conducted at the elementary level (i.e., 2nd and 4th grades) of two demographically similar school districts, District C and District D. Sixteen teachers from 9 schools participated in the mathematics portion of the study, and 14 of these teachers turned in assignments with samples of student work. Teachers who participated in the study had been teaching for an average of nine years, and had been at their school an average of three years. The total number of students from the participating teachers' classes was 336. The classes contained a diverse population of students (16% African American, 8% Asian, 63% Latino, 11% white, 2% other), 19% of whom were English language learners.

Two very important differences exist between District C and District D. First, the administrators and teachers in District C had a long-standing relationship with the Institute for Learning (IFL), which had provided the district with regular professional development sessions involving consistent and sustained efforts to implement the Principles of Learning into their schools and classrooms. On the other hand, the administrators and teachers in District D were in the initial phases of their partnership with the IFL and of their implementation of the Principles of Learning.

Districts at each end of the professional development spectrum, with regard to their history with the IFL and opportunities to implement the Principles of Learning, were purposefully chosen as a way of discerning whether the IQA rubrics were able to uncover the differences in instructional practices that they were designed to measure. Second, District C used an elementary mathematics curriculum designed to engage students in learning mathematics with understanding (as described earlier in this report) that contained a predominance of mathematical tasks involving thinking, reasoning, and sense-making (i.e., high-level tasks). The elementary mathematics curriculum in District D had a skill-based focus of improving students' accurate and efficient use of mathematical procedures and memorization of mathematical facts (i.e., low-level mathematical tasks). These differences provide a lens through which to interpret the descriptive statistics for each district in the analysis section.

Procedures

For the pilot study, six raters were recruited from graduate schools of education in universities near the districts participating in the study. Raters underwent a 2.5-day training program designed and administered by IQA developers. The training included reading about the research base upon which the AR-Math rubrics are based; a brief overview of the content of the primary-grades mathematics curriculum based on NCTM Standards (2000) and a selection of popularly used curricular materials (both traditional and reform-oriented); practice rating instructional tasks selected from a sample of primary grades mathematics curricula; practice rating tasks implementation and student discussion based on video and written episodes of instruction; and practice identifying and rating teachers' expectations based on video clips of mathematics instruction. The raters were not told why the districts were selected for the study, and were unfamiliar with the IQA prior to training.

The lesson observations occurred over a two-week period starting with District C and followed by District D. Two trained raters, accompanied by an IQA staff member, observed one full lesson (45-50 minutes) for each teacher and interviewed up to 4 students per classroom. During the observation, raters made detailed field notes and used them to provide evidence to justify their scores. Following the conclusion of the lesson, each rater scored the lesson independently, using the IQA rubrics of Accountable Talk (AT), Clear Expectations (CE), and Academic Rigor-Mathematics (AR-Math); again, just the AR-Math rubrics and their results are

discussed here. The raters then debriefed their individual scores, and in instances of disagreement, reached a consensus score. Raters' individual scores are used for the analyses presented in this report.

Each teacher also submitted four mathematics assignments with samples of students' work. Our collection of student work is based on research by Matsumura and colleagues (Matsumura, et al., 2002; Clare & Aschbacher, 2001), which determined that student work collections consisting of four samples each (two medium quality and two high quality) and rated by two raters would yield a generalizability co-efficient high enough (i.e., G > .80) to use assignments and student work as valid indicators of classroom practice. For each assignment, teachers filled out a two-page cover sheet describing the assignment task, their assessment criteria for grading student work and how they shared these criteria with students. The teachers' responses on the cover sheet were used as source of evidence to rate the rubrics for academic rigor in the assignments. To prepare for rating assignments, the raters attended a one-day rater-training program and independently scored a total of 55 assignments that were randomly ordered. A two-day assignments rating took place three weeks after the lesson observations when the assignment collection was complete. Similar to the lesson observations, the raters debriefed their scores for the assignment ratings and reached a consensus score when disagreements occurred.

Analysis

In order to assess the quality of our rater-training program and rubrics, rater reliability was first examined by computing several different measures. Percent of exact agreement between the two raters' independent scores was first computed. Cohen's kappa coefficients were calculated to investigate the level of agreement between the two raters on each dimension when controlled for chance agreement. Correlations were also computed to measure the strength of agreement between the rater pair. Second, descriptive statistics were used to characterize the lesson observations and assignments, and pairwise t-tests were used to make comparisons between school districts in each of the four dimensions for the AR-Math Lesson Observation and Assignment rubrics. Third, a generalizability study (G-study) was conducted to investigate whether the design based on two raters and the collection of four assignments from teachers yielded a stable estimate of the overall quality of teachers' instructional practices. This analysis was expected to provide information on the sources of variation as well. Finally, correlations were computed at the

teacher-level to investigate the interrelationship within the observed lesson ratings and within the assignment ratings. Descriptive statistics and correlation analyses were conducted based on the consensus scores between the two raters.

Results

Reliability

Reliability tests were conducted to compare the agreement of the two raters' initial, independent scores in each dimension. The percent agreement between raters was calculated on the overall mathematics rubrics and on each dimension within the AR-Math rubrics for Lesson Observations and for Assignments.

For Lesson Observations, results indicate a poor level of exact agreement between the two raters on the overall math rubrics (50.0%), though 1-point agreement was excellent (95.2%). Table 1 displays the results of rater reliability for individual dimensions within the AR-Math rubric. The reliability for Lesson Observations ranged from poor to fair, with percentages of exact agreement between 37.5% for the *Potential of the Task* rubric (AR1) and 70.0% for the *Discussion* rubric (AR3). The Kappa coefficients ranged from fair to moderate, being far from a satisfactory level of .70 (Gardner, 1995). The correlation between raters was also poor (r = .34 to r = .42) for each rubric except *Discussion* (r = .72). The Assignment ratings, which occurred after the Lesson Observations, exhibited higher reliability levels than the Lesson Observation ratings overall (63.5% for exact agreement; 97.4% for 1-point agreement) and in each of the AR-Math rubrics (ranging from 60.0% to 67.3%). The Kappa coefficients and correlation coefficients indicated that there was a moderate level of agreement between the two raters.

Reliability was also analyzed by collapsing each rubric to a 3-point scale. Rater agreement after regrouping the 4-point score scales increased substantially by grouping levels 3 and 4 together and by grouping levels 2 and 3 together (see Table 2). By collapsing score levels 3 and 4, each dimension also increased its percentage of exact agreement between raters, with the *Potential of the Task* (AR1) increasing considerably compared to the other rubrics (see Table 3). The collapsed scales were intended to inform future rater-training efforts by identifying whether inconsistencies in raters' scores were the result of confusion between specific score levels overall and within each AR-Math dimension.

Table 1

Inter-rater reliability for AR-Math Rubrics for Lesson Observation and Assignment scores

AR-Math Rubrics	Lesson Observation (N = 16 teachers)			Assignments (N = 55 assignments)		
	% of exact agreement	Kappa	Spearman r	% of exact agreement	Kappa	Spearman r
AR1: Potential	37.5	-	.39	65.5	.51	.73
AR2: Implementation	50.0	.27	.42	60.0	.43	.72
AR3: Discussion*	70.0	.54	.72	67.3	.53	.74
AR4: Expectations	53.3	.33	.34	63.6	.43	.68

* AR3 in the Assignment ratings indicates the Rigor of students' written response dimension.

Table 2

Rater Reliability of Lesson Scores after Regrouping Score Scales for Overall AR Rubrics (N = 16 teachers)

	% of exact agreement	Kappa
4 point scale (1 - 4)	50.0	.29
3 point scale (1, 2, and ¾)	64.6	.36
3 point scale (1, 2/3, and 4)	63.0	.36
3 point scale (1/2, 3, and 4)	56.1	.34

Table 3

Exact Agreement of Lesson Scores after Regrouping Score Scales for AR-Math Rubrics (1, 2, and 3 /4) (N = 16 teachers)

AR Rubrics	% of exact agreement
AR1: Potential	60.0
AR2: Implementation	60.0
AR3: Discussion	66.7
AR4: Expectations	66.7

Quality of Instruction through Lesson Observations and Assignments

Descriptive statistics were computed to characterize students' opportunities to learn mathematics with understanding with respect to each dimension of the AR-Math rubrics. These measures also allowed for comparisons between the two school districts in the study.

For lesson observations (see Table 4), students in both districts were provided with similar levels of tasks (AR1), but the level at which these tasks were implemented (AR2) was higher in District C than in District D. The low variance in AR2 for District D indicates that instruction is typically characterized by procedures without connection to meaning or understanding. Teachers' expectations also differed significantly, with teachers in District D having a lower level of expectations than their counterparts in District C. The score distribution in Appendices A and B for the Lesson Observation and Assignment rubrics provide another portrayal of the differences in students' opportunities to learn in each of the districts.

Table 4

Summary of Descriptive Statistics and t-test of Scores on AR-Math Lesson Observation Rubrics by Districts

AR Rubrics	District C *Mean (SD)* (n = 8)	District D *Mean (SD)* (n = 8)	Mean Difference	t
AR1: Potential	2.75 (.46)	2.50 (.76)	0.25	0.798
AR2: Implementation	2.63 (.52)	2.13 (.35)	0.50	2.256*
AR3: Discussion	2.50 (.84)	1.80 (.84)	0.70	1.382
AR4: Expectations	2.88 (.64)	2.00 (.93)	0.88	2.198*

*$p < .05$

For the Assignment rubrics (see Table 5), all four dimensions were significantly different in favor of District C. Scores for District C indicate that students are frequently provided with opportunities to engage in high-level tasks (mean for AR1 > 3.0). These tasks are often implemented in ways that maintain the high-level cognitive demands (mean for AR2 = 2.64) and that provide evidence of high-level cognitive demands in students' written responses (AR3 = 2.67). Teacher's expectations almost always consist of high-level requirements for students' work (AR4 > 3.0).

In contrast, scores for District D in each dimension lie below the demarcation line between high- and low-level cognitive demands, indicating that mathematics instruction and learning in District D is not typically characterized by understanding, sense-making, or use of a variety of representations or problem-solving strategies. Rather, with the mean score for each dimension falling under a 2.0, students' opportunities for learning mathematics tend to emphasize prescribed procedures that are not connected to meaning and understanding and/or memorization. An argument can be made that, when taking the variance into consideration, the mean scores reflect a mixture of tasks at each level; however, even when considering the range of scores that fall within 1 standard deviation of the mean, students in District D are almost never provided with tasks that have the potential to be a 4 (AR1) and rarely engage with tasks (AR2), provide responses (AR3), or are given expectations (AR4) with a high level of cognitive demand (i.e., at or above a score of 3).

Table 5

Summary of Descriptive Statistics of Scores and t-test on AR-Math Assignment Rubrics by Districts

AR Rubrics	District C *Mean (SD)* (*n* = 27 assignments)	District D *Mean (SD)* (*n* = 28 assignments)	Mean Difference	*t*
AR1: Potential	3.15 (.53)	1.93 (.72)	1.22	7.138*
AR2: Implementation	2.63 (.79)	1.61 (.69)	1.02	5.127*
AR3: Responses	2.67 (.78)	1.50 (.79)	1.17	5.482*
AR4: Expectations	3.07 (.39)	1.96 (.58)	1.15	8.384*

*$p < .05$

Generalizability (G) Study

A G-study was conducted to determine whether our design for rating assignment collections yielded a stable estimate of the quality of classroom practice. Results indicated that our design based on two raters and four sets of assignments per teacher yielded an excellent G coefficient of .91 (.80 and above is considered to be good). As shown in Table 6, 55.2% of the variance was explained by the variation between teachers, indicating a considerable amount of systematic variability between teachers in their instructional practices. Ten percent of the variance was explained by the interaction between the teacher and the assignment type,

suggesting that the teachers submitted different types of assignments. The variance component for Rubrics accounts for only 2% of the total variance. This result suggests that the Rubrics measured a coherent construct, that is, the rigor of the mathematical content. The overall results of the G-study lend support to the contention that students' assignments can be used as indicators of classroom practice (Clare & Aschbacher, 2001).

Table 6

Estimates of Variance Components for the Mathematics Assignments (N = 14 teachers, 55 assignments)

Source of Variation	Estimated Variance Component*	Percentage of Total Variance
Teacher	0.530	55.2
Rater	0.000	0.0
Assignment Type	0.000	0.0
Rubric	0.023	2.4
Teacher x Rater	0.000	0.0
Teacher x Assignment Type	0.097	10.1
Teacher x Rubric	0.013	1.4
Rater x Assignment Type	0.010	1.0
Rater x Rubric	0.000	0.0
Assignment Type x Rubric	0.006	0.6
Teacher x Rater x Assignment Type	0.053	5.5
Teacher x Assignment Type x Rubric	0.013	1.4
Rater x Assignment Type x Rubric	0.000	0.0
Teacher x Rater x Assignment Type x Rubric, Error	0.215	22.4

*Negative variance component was set to zero.

Relationships among Rubric Dimensions

The results of correlation analyses indicate that all four AR-Math dimensions were significantly correlated within Lesson Observation rubrics and Assignment rubrics. These results are provided in Tables 7 and 8. Particularly, the *Potential of the Task* and the *Teacher's Expectations* are highly correlated for the mathematics rubrics.

Table 7

Inter-correlation within Lesson Observation Scores on AR-Math Rubrics (N = 16 teachers)

	Lesson Observation			
	AR1	AR2	AR3	AR4
AR1: Potential	-	.71**	.82**	.89**
AR2: Implementation		-	.68*	.68**
AR3: Discussion			-	.82**
AR4: Expectations				-

*p < .05. ** p < .01

Table 8

Inter-correlation within Assignment Scores on AR: Math Rubrics (N=14 teachers)

	Assignment			
	AR1	AR2	AR3	AR4
AR1: Potential	-	.81*	.80*	.82*
AR2: Implementation		-	.87*	.74*
AR3: Responses			-	.70*
AR4: Expectations				-

*p < .01

Discussion

Reliability

Overall, exact-point reliability between rater-pairs ranged from poor to moderate, and 1-point reliability was excellent. A great deal of disagreement about *Potential of the Task* rubric surfaced during the rater debriefing for the mathematics lesson observations. Hence, low reliability in that rubric was not a surprising result, and suggests that more training is required for rating the potential of the task. This contention is supported by the fact that reliability increased over time (i.e., from District C to District D; from lesson observations to assignments; from practice assignment ratings to actual assignment ratings), indicating a general trend that more experience leads to greater reliability.

Reliability results on the collapsed scales were intended to inform future rater-training efforts by identifying whether inconsistencies in raters' scores resulted from confusion between specific score levels overall and within each AR-Math dimension. One source of confusion appeared to lay between the score levels of 3 and 4, as evidenced by the improved reliability when levels 3 and 4 were combined. In each rubric, levels 3 and 4 are both representative of high-level cognitive demands, with level 4 requiring explicit mathematical connections or reasoning. Raters' difficulty in distinguishing between levels 3 and 4 indicates that more training is needed specifically in determining what constitutes (or does not constitute) evidence of explicit high-level cognitive processes in each rubric. Similarly, reliability improved when score levels 2 and 3 were combined, and this finding also has specific implications for rater-training. Recall the demarcation line between high and low-level cognitive demands between the scores of 2 and 3 in each dimension. Failure to distinguish between a 2 and a 3 would indicate that raters had difficulty differentiating between high- and low-level cognitive demands. The main difference between levels 2 and 3 is whether the mathematical task is connected to meaning, understanding, and sense-making, again indicating the need for additional rater training in what constitutes evidence of connections to meaning and understanding in mathematics. This training should also provide raters with opportunities to generalize characteristics of high vs. low-level cognitive demands in each dimension of the AR-Math rubrics.

In summary, these results suggest that it was hard for raters to distinguish between score levels of 3 and 4, as well as 2 and 3, in each dimension of the math rubrics. The finding that reliability improved with time and experience is encouraging, and appears to indicate that more training is needed, especially for the *Potential of the Task* dimension. We anticipate that reliability results will increase with expert raters highly knowledgeable in mathematics education or with the IQA rubrics. Expert rater-pairs would enter classrooms with consistent ideas of what the rubric levels look like overall and within each dimension, specifically with respect to what constitutes evidence of mathematical connections (i.e., differentiating between score levels 3-4 and 2-3). This distinction appeared to be particularly difficult for newly-trained raters, none of which had backgrounds in mathematics education or any familiarity with the IQA prior to rater-training.

Validity

The analysis of students' opportunities to learn mathematics in the two districts suggests that students in each district get to engage with high-level tasks during mathematics instruction, and that this occurs more frequently in District C. Note that in both districts, lesson tasks tended to be implemented at lower levels of cognitive demand than the potential level of cognitive demand of the task. This finding supports the contention that maintaining high-level cognitive demands is a challenging endeavor for mathematics teachers (Hiebert, et al., 2003).

Significant differences between District C and District D can be the result of several factors. First, District C had a long-standing professional development partnership with the Institute for Learning and was considered a "high-implementation" district in regards to adoption and enactment of the Principles of Learning. District D was in the beginning stages of professional development and implementation of the Principles of Learning. Because the IQA rubrics are designed to assess quality instruction through the assessment of four Principles of Learning (Accountable Talk, Academic Rigor, Clear Expectations, and Self-Management of Learning), it follows that District C would naturally score higher. Teachers in District C had substantially more opportunities to learn to enact instruction consistent with the ideals upon which the AR-Math rubrics were based. An interesting use of the IQA toolkit would be to reassess District D at some point in the future to determine their areas of growth. Second, the mathematics curriculum in District C contained a predominance of high-level tasks (i.e., a 3 or 4 in *Potential of the Task*) and provided support to teachers in implementing these tasks in ways that maintained the high-level cognitive demands (i.e., a score of 3 or 4 in *Implementation*). Hence, teachers in District C had more access to high-level tasks and more support to enact tasks at a high-level, as well.

Relationships between Rubric Dimensions

All AR-Math rubric dimensions were significantly correlated. This correlation may indicate a redundancy in rubric dimensions or, conversely, may be a desired outcome of the rubrics and of mathematics instructional programs. Determining whether certain dimensions are redundant, and thus can be eliminated, might differ based on statistical relevance vs. practical relevance in informing mathematics instruction at the school or teacher level. For example, students' discussion is a subset of Task Implementation. However, we contend that the rigor of the

discussion is important enough to tease out and assess independently from the overall lesson. Other instructional questions may arise if certain dimensions were not correlated. Would such consistency (or lack thereof) provide important information about students' opportunities to learn mathematics in teachers' classrooms and/or in school mathematics instructional programs? For instance, what would be the instructional implications if teachers' expectations were not consistent with the potential of the task? Such answers are currently beyond the scope of this small scale pilot study, but the success of the pilot in raising these issues as avenues for future investigation is invaluable.

Answering the Overarching Question: Is the IQA Toolkit an Effective Tool for Assessing Academic Rigor in School Mathematics Programs?

We contend that, based on the above results, the IQA toolkit appears to be an effective tool for evaluating school mathematics programs. As identified by the descriptive statistics, the rubrics teased out important differences in students' opportunities to learn mathematics in each district. Furthermore, results at the district level were very indicative of the nature and extent of reform efforts in District C as compared to District D. Differences in the two districts identified by the descriptive statistics seem consistent with the high inter-correlations in rubric dimensions: individual teachers tended to score similarly on all dimensions, with teachers in District C tending to have consistently high scores and teachers from District D tending to have consistently lower scores.

One similar feature between the two districts is that tasks tended to be implemented in lessons and enacted in students' assignments at lower levels of cognitive demand than what the task had the potential to offer. This finding suggests the need for professional development specifically designed to assist teachers in maintaining high-level cognitive demands through an instructional episode and in fostering the development of high-level cognitive processes in students' work. At the teacher level, the rubrics identified differences between teachers in the level of assignments provided to students. Synthesizing this result with the high correlation between the potential of the assignment tasks and teacher's expectations might indicate that individual teachers tend to give assignments of consistent with their level of expectations for students' learning. Hence, raising teachers' expectations can in turn generate increases in students' opportunities to engage with high-level tasks—thereby increasing students' opportunities to learn mathematics with understanding.

Summary and Implications

The Strengths (and Weaknesses) of the IQA Academic Rigor in Mathematics Rubrics

The IQA AR-Math rubrics have a very specific theoretical basis. Constructivist learning environments—and NCTM's vision of teaching and learning mathematics constitute the foundation of our work. Furthermore, we have built our analysis on specific aspects of mathematics instruction known to influence students' opportunities to learn mathematics (the level of cognitive demand of instructional tasks, task implementation, the discussion, and the teachers' expectations), and we have based our rubrics on observation tools and methods of analyzing mathematics instruction that are capable of teasing out important difference in students' opportunities to learn (e.g., QUASAR, TIMSS). Through our specific focus on tasks, task implementation, discussions, and teachers' expectations through the lens of the level of cognitive demands, we have attempted to focus on observable aspects of mathematics instruction that influence the rigor of a mathematics lesson and require low inference on the part of the rater. Furthermore, we anticipate that the descriptors for the score levels in the AR-Math rubrics are specific enough to provide a clear picture of what instruction looked like in that particular classroom, school, or district. In this way, we hope that the AR-Math rubrics can provide a path toward instructional change in a direction that is empirically and theoretically justified as leading to improved opportunities for students to learn mathematics.

The value of a pilot study can be determined by its ability to identify areas in need of improvement. In this sense, we can consider our initial pilot of the IQA rubrics reported herein to be quite successful. Low inter-rater reliability indicates that many of the rubric dimensions were not as evident or low-inference as we would have hoped. Given that the inter-rater reliability results (1) are based on newly trained, non-expert raters and (2) improved with increased exposure to and experience with the rubrics, we are optimistic that reliability scores would increase substantially with expert raters and with enhancements to our rater-training program. The pilot also provided us with a direction for improving rater-training. Specifically, raters need more practice (1) rating the potential of the task, (2) identifying evidence of mathematical connections and reasoning, and (3) identifying general differences between high- and low-level cognitive demands in mathematics. We were pleased that the results of the G-studies confirmed earlier research on the validity of using students' assignments as indicators of quality instruction in

mathematics. Overall, we contend that the Academic Rigor in Mathematics rubrics of the IQA Toolkit identify important differences in students' opportunities to learn mathematics. Of course, our conclusions at this point are tentative, based on a small-scale pilot study, and will require further research. Such questions that might be addressed include: Will reliability improve amongst raters highly knowledgeable about mathematics education or about the IQA rubrics? Do students in classrooms rated highly by the AR-Math rubrics exhibit high levels of student achievement?

A validity study that incorporates student achievement data is currently in development. We hope to add to the knowledge base of what aspects of quality instruction appear to provide increased opportunities for students' learning. We also intend to develop an internal version of the IQA that can be used in professional development for teachers of mathematic. This paper closes by describing how the IQA AR-Math rubrics might serve as a valuable professional development resource.

Implications for Mathematics Teacher Development: Selecting and Implementing High-Level Tasks

This report will close by offering a suggestion for the future use of the IQA toolkit: the professional development of mathematics teachers. In choosing to focus on the mathematical tasks with which students engage during mathematics instruction, the IQA rubrics draw on research and theories ascertaining the importance of exposing students to high-level mathematical tasks (Stein & Lane, 1996; Doyle, 1988; Hiebert & Wearne, 1983). One of the most critical responsibilities of mathematics teachers is to provide students with tasks that encourage mathematical thinking, reasoning, and problem-solving (Doyle, 1988; Hiebert et al., 1997). The need for mathematics teachers to determine what constitutes a high-level task, to assess whether a task can provide the types of learning opportunities that promote students' understanding, is thus of prime importance. The IQA AR-Math rubrics can help teachers to analyze the cognitive demands of mathematical tasks, differentiate between tasks with high- and low-level cognitive demands, and identify features of tasks that promote student engagement with high-level cognitive demands.

In using the IQA AR-Math rubrics as a professional development tool, teachers will also be exposed to a framework for analyzing the cognitive demands of mathematical tasks throughout an instructional episode. Selecting high-level tasks is the first step in improving students' opportunities to learn mathematics with

understanding; implementing these tasks in ways that maintain students' opportunities to engage in high-level cognitive processes is the second step. The AR-Math rubrics and the lesson observation checklist can help teachers identify important factors in maintaining high-level cognitive demands throughout an instructional episode.

Exposing teachers to the AR-Math rubrics is hypothesized to serve as a catalyst for instructional change by having a "teaching to the test" effect—teachers will change their instructional practices to reflect the nature and content of the dimensions on which they are being assessed. Hence, the IQA AR-Math rubrics can potentially serve as a professional development tool to engage teachers of mathematics in identifying high-level instructional tasks, in implementing these tasks in ways that maintain the high-level cognitive demands, in orchestrating mathematical discussions that provide students with opportunities to make mathematical connections, and in having high-level expectations for their students. In this way, the IQA toolkit serves not only as a means of assessing quality instruction, but also as a tool for promoting quality instruction in school systems and in classrooms by improving students' opportunities to learn mathematics with understanding.

References

Ben-Chaim, D., Fey, J. T., Fitzgerald, W. M., Benedetto, C., & Miller, J. (1998). Proportional reasoning among 7th grade students with different curricular experiences. *Educational Studies in Mathematics, 36*(3), 247-273.

Borko, H., Stecher, B., Alonzo, A., Moncure, S., & McClam, S. (2003). *Artifact packages for measuring instructional practice: A pilot study* (CSE Tech. Rep. No. 615). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chazan, D., & Ball, D. L. (2001). Beyond being told not to tell. *For the Learning of Mathematics, 19*(2), 2-10.

Clare, L., & Aschbacher, P. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment, 7*(1), 39-59.

Clarke, D. M. (1997). The changing role of the mathematics teacher. *Journal for Research in Mathematics Education, 28*(3), 278-308.

Cobb, P., Boufi, A., McClain, K., & Whitenack, J. (1997). Reflective discourse and collective reflection. *Journal for Research in Mathematics Education, 28*(3), 258-277.

Doyle, W. (1988). Work in mathematics classes: The context of students' thinking during instruction. *Educational Psychologist, 23*(2), 167-180.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53,* 159-199.

Fuson, K. C., Carroll, W. M., & Druek, J. V. (2000). Achievement results for second and third graders using standards-based curriculum. *Everyday Mathematics. Journal for Research in Mathematics Education, 31*(3), 277-295.

Gardener, W. (1995). On the reliability of sequential data: Measurement, meaning, and correction. In J. M. Gottman (Ed.), *The analysis of change.* Mahwah, NJ: Erlbaum.

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*(5), 524-549.

Hiebert , J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning.* New York: Macmillian.

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K. C., Wearne, D., Murray, H., Olivier, A., & Human, P. (1997). *Making sense: Teaching and learning mathematics with understanding*. Portsmouth, NH: Heinemann.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K., Hollingsworth, H., Jacobs, J., Chui, A. M., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E. A., Etterbeek, W., Manaster, C., & 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.

Hiebert, J., & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal, 30*(2), 393-425.

Huntley, M. A., Rasmussen, C. L., Villarubi, R. S., Sangtong, J., & Fey, J. T. (2000). Effects of *Standards*-based mathematics education: A study of the *Core-Plus Mathematics* project algebra and function strand. *Journal for Research in Mathematics Education, 31*(3), 328-361.

Institute for Learning (2002). *Principles of Learning*. Overview available at http://www.instituteforlearning.org/pol3.html. University of Pittsburgh. Pittsburgh, PA: Author.

Junker, B., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., Weisberg, Y., & Resnick, L. (2004, April). *Overview of the Instructional Quality Assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Lampert, M. (2001). *Teaching problems and the problems of teaching.* New Haven, CT: Yale University Press.

Lesh, R. A., Post, T. R., & Behr, M. J. (1987). Representations and translations among representations in mathematics learning and problem solving. In C. Janvier (Ed.), *Problems of representation in the learning of mathematics.* Mahwah, NJ: Lawrence Erlbaum.

Matsumura, L. C., Garnier, H. E., Pascal, J., & Valdes, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement (CSE Tech. Rep. No. 582). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Matsumura, L. C., Wolf, M. K., Crosson, A. (2004). *Assessing the Quality of Reading Comprehension Assignments and Student Work*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional Standards for Teaching Mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: Author.

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: Cambridge University Press.

Putnam, R. T, Lampert, M., & Peterson, P. L. (1990). In C. B. Cazden, (Ed.), *Review of research in education* (Vol. 16, pp. 57-150). Washington, DC: American Educational Research Association.

Remillard, J. (1999). Curriculum materials in mathematics education reform: A framework for examining teachers' curriculum development. *Curriculum Inquiry, 29*(3).

Resnick, L. B., & Hall, M. W. (1998). Learning organizations for sustainable education reform. *Daedalus, 127*, 89-118.

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*(4), 368-398.

Romanagno, L. (1994). *Wrestling with change: The dilemmas of teaching real mathematics.* Portsmouth, NH: Heinemann.

Romberg, T. A., & Carpenter, T. P. (1986). Research on teaching and learning mathematics: Two disciplines of scientific inquiry. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 850-873). American Educational Research Association.

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*(1), 74-95.

Schoen, H. L., Fey, J. T., Hirsch, C. R., & Coxford, A. F. (1999). Issues and opinions in the math wars. *Phi Delta Kappan, 80*(6), 444-53.

Schoenfeld, A. H. (2002). Making mathematics work for all children: issues of standards, testing, and equity. *Educational Researcher, 31*(1), 13-25.

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*(January 1996), 476-521.

Smith, M. S. (1995). A road to change. Doctoral dissertation, University of Pittsburgh. (UMI No. 9614231).

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*(2), 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.

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*(1), 58-84.

United States Department of Education. (1999). Exemplary and promising mathematics programs: Expert panel report. Available at www.enc.org/professional/federalresources/exemplary/promising/

Voigt, J. (1994). Negotiation of mathematical meaning and learning mathematics. *Educational Studies in Mathematics, 26,* 275-298.

Weiss, I. R., & Pasley, J.P. (2004). What is high quality instruction? *Educational Leadership, 61*(5), 24-28.

Wolf, M. K., Crosson, A., & Resnick, L. B. (2004). Classroom talk for rigorous reading comprehension instruction. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Appendix A

Score Distributions of AR-Math Lesson Observation Rubrics by District

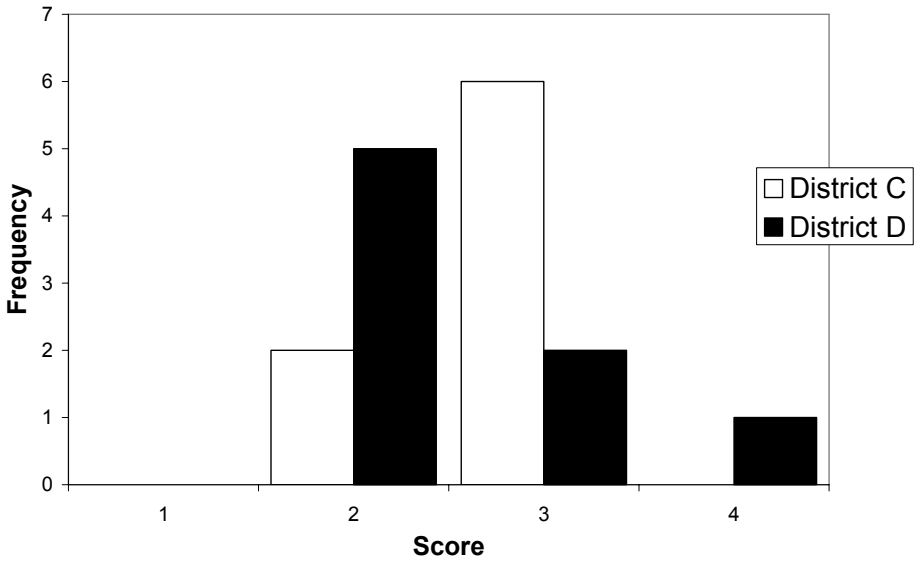

Figure 1. Score distribution on AR1 (*Potential of Task*) of the two districts.

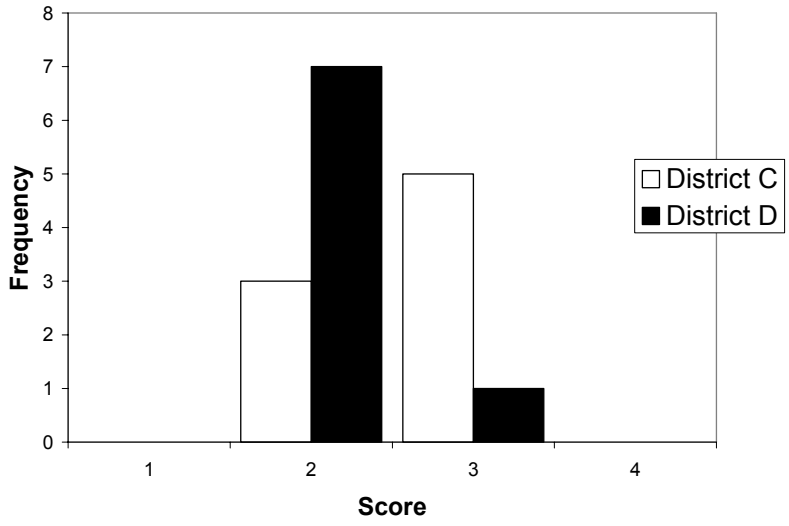

Figure 2. Score distribution on AR2 (*Implementation of Task*) of the two districts.

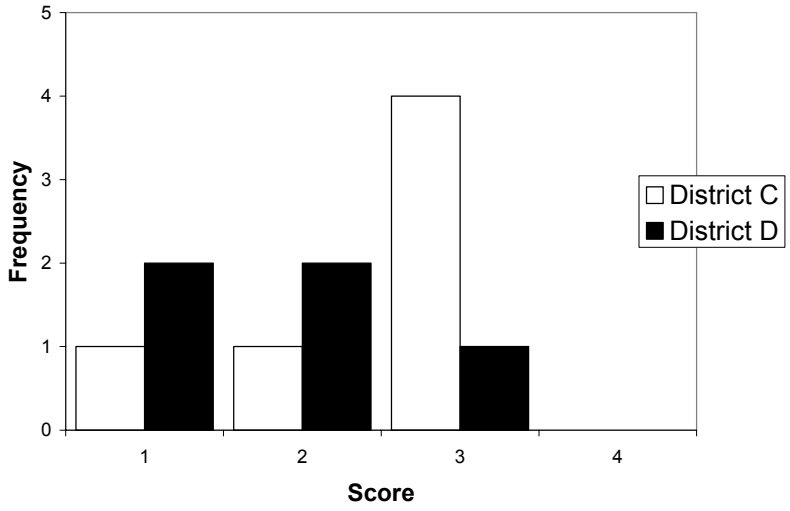

Figure 3. Score distribution on AR3 (*Discussion*) of the two districts.

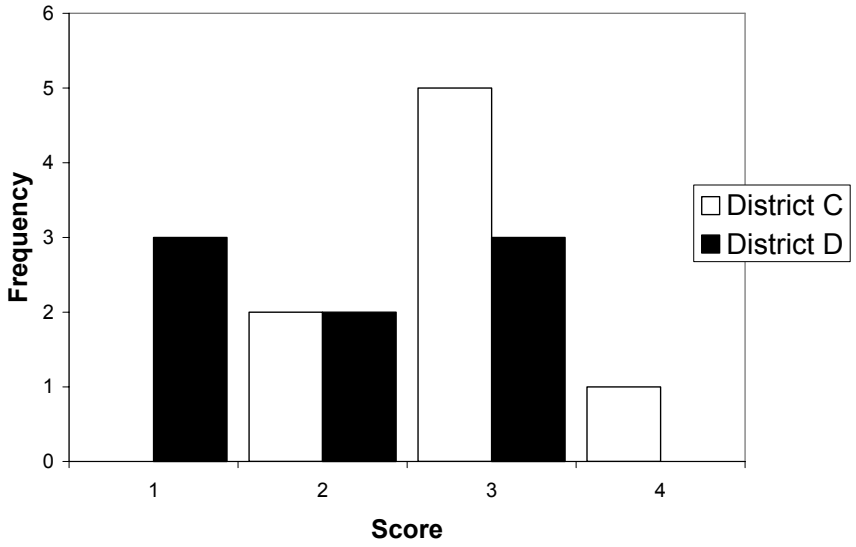

Figure 4. Score distribution on AR4 (*Rigor of Expectations*) of the two districts.

Appendix B

Score Distributions of AR-Math Assignment Rubrics by District

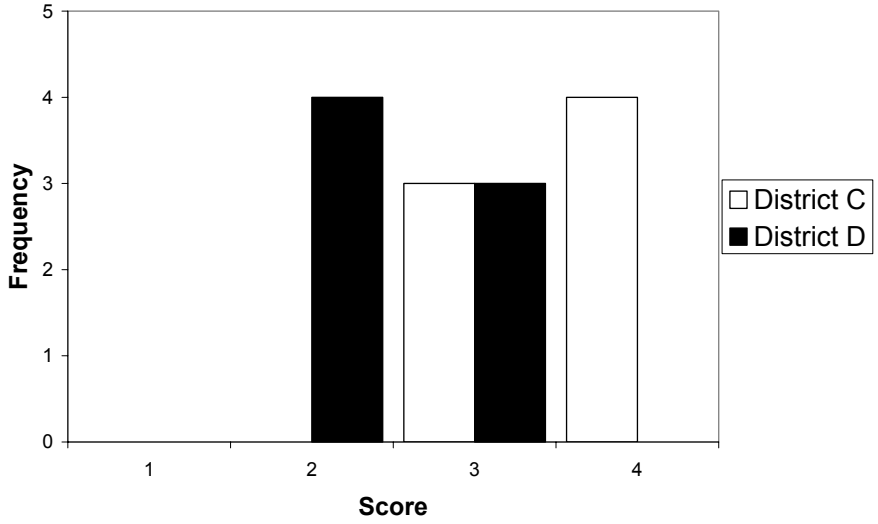

Figure 5. Score distribution on AR1 (*Potential of Task*) of the two districts.

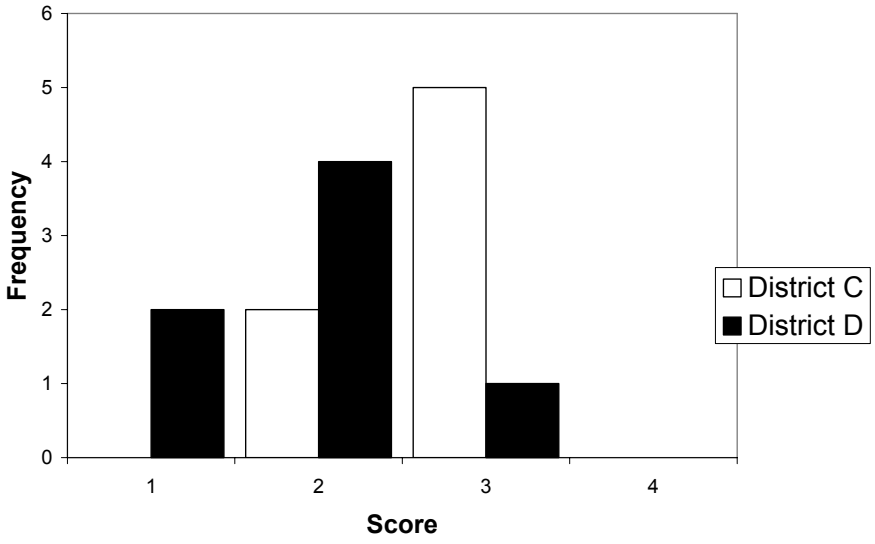

Figure 6. Score distribution on AR2 (*Implementation of Task*) of the two districts.

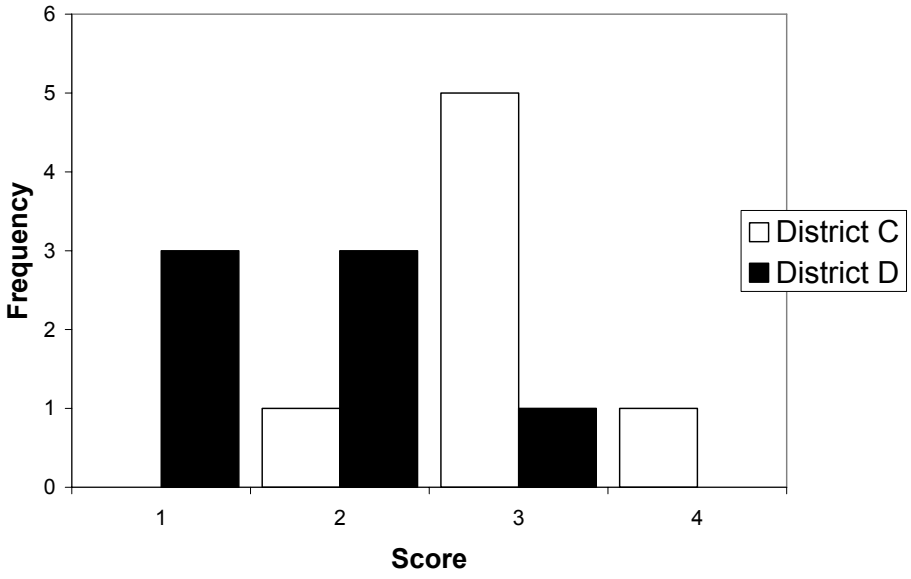

Figure 7. Score distribution on AR3 (*Response*) of the two districts.

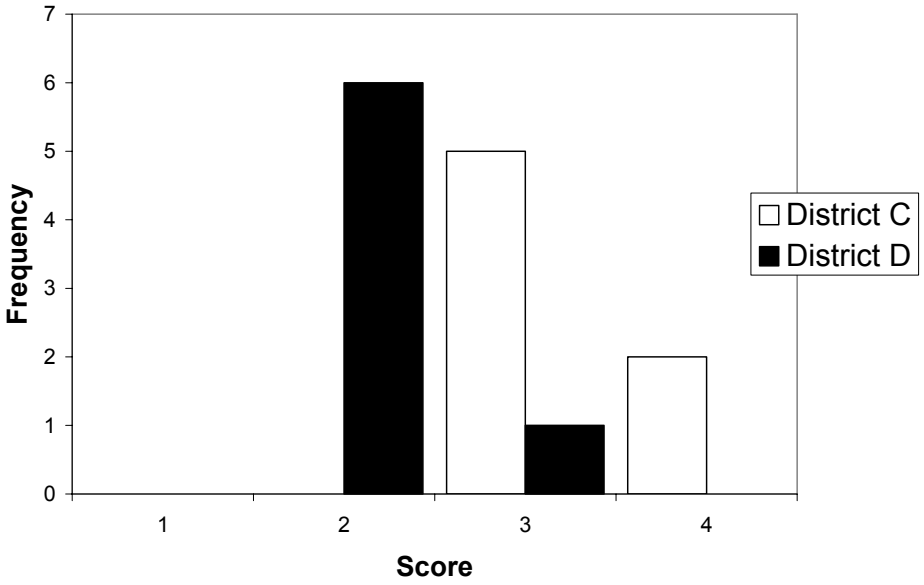

Figure 8. Score distribution on AR4 (*Rigor of Expectations*) of the two districts.

Appendix C: 2003 Draft Observation and Assignment Rubrics for Mathematics

For revised 2005 version of the rubrics, please contact:

Dr. Lindsay Clare Matsumura, lclare@pitt.edu

Dr. Brian Junker, brian@stat.cmu.edu

Accountable Talk Observation Rubrics, 2003

Consider talk from the whole-group discussion only.

1. How effectively did the lesson-talk build <u>Accountability to the Learning Community</u>?

Low-inference dimensions, to be rated after observing all teacher-facilitated discussions of the lesson:

A. Participation

Was there widespread participation in teacher-facilitated discussion?

4	Over 50% of the students participated consistently throughout the discussion.
3	25-50% of the students participated consistently in the discussion OR over 50% of the students participated minimally.
2	25-50% of the students participated minimally in the discussion (i.e, they contributed only once).
1	Less than 25% of the students participated in the discussion.
N/A	Reason:

B. Linking contributions

Did speakers' contributions link to and build on each other? (i.e., Was there "local coherence" during the discussion?)

4	At at least 3 points during the discussion, the teacher/student explicitly connects speakers' contributions and shows how ideas/positions shared during the discussion relate to each other.
3	At 1-2 points during the discussion, the teacher / student links speakers' contributions to each other and shows how ideas/positions relate to each other.
2	At one or more points during the discussion, the teacher / student links speakers' contributions to each other, but does not show how ideas/positions relate to each other.
1	Teacher / student does not make any effort to link speakers' contributions.
N/A	Reason:

1. Teacher contributions 4 ___ 3 ___ 2___ 1___
2. Student contributions 4 ___ 3 ___ 2___ 1___

2. How effectively did the lesson-talk build <u>Accountability to Knowledge</u>?

Asking: Were contributors asked to support their contributions with evidence?

4	There are 3 or more efforts to ask students to provide evidence for their contributions, including questions that seemed academically relevant.
3	There are 1-2 efforts to ask students to provide evidence for their contributions that seemed academically relevant.
2	There are one or more superficial, trivial efforts, or formulaic efforts to ask students to provide evidence for their contributions.
1	There are no efforts to ask students to provide evidence for their contributions.
N/A	Reason:

Providing: Did contributors support their contributions with evidence? (This evidence must be appropriate to the content area—i.e., evidence from the text; citing an example, referring to prior classroom experience.)

4	At at least 3 points, speakers provide accurate and appropriate evidence for their claims, including frequent references to the text or prior classroom experience.
3	At 1-2 points, speakers provide accurate and appropriate evidence for their claims, including references to the text or prior classroom experience.
2	In general, what little evidence is offered to back up claims is inaccurate, incomplete, or vague.
1	Speakers do not back up their claims.
N/A	Reason:

3. How effectively did the lesson-talk build <u>Accountability to Rigorous Thinking</u>?

Asking: Were speakers asked to explain their thinking during the lesson?

4	There are 3 or more efforts to ask students to explain their reasoning, including questions that seemed academically relevant.
3	There are 1-2 efforts to ask students to explain their reasoning that seemed academically relevant.
2	There is at least one superficial, trivial, or formulaic efforts to ask students to explain their reasoning.
1	There were no efforts to ask students to explain their thinking.
N/A	Reason:

Providing: Did contributors explain their thinking during the lesson?

4	There are 3 or more examples of speakers explaining their thinking, using reasoning in ways appropriate to the discipline.
3	There are 1-2 examples of speakers explaining their thinking, using reasoning in ways appropriate to the discipline.
2	In general, what little attempt to explain reasoning is vague or inappropriate.
1	Speakers do not explain the reasoning behind their claims.
N/A	Reason:

A. Potential of the Task	
4	The task has the potential to engage students in "doing mathematics" or "procedures with connections" : • using complex and non-algorithmic thinking (i.e., there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instructions, or a worked-out example); • exploring and understanding the nature of mathematical concepts, procedures, and/or relationships. The task may require students to: • solve a genuine, challenging problem; • develop an understanding for why formulas or procedures work; • apply a broad general procedure that remains closely connected to mathematical concepts; • identify patterns and form generalizations based on these patterns; • make conjectures and support conclusions with mathematical evidence; • make connections between representations, strategies, or mathematical concepts and procedures.
3	The task has the potential to engage students in complex thinking or in creating meaning for mathematical concepts, procedures, and/or relationships. However, the task does not warrant a "4" because: • students engage in problem solving, but the mathematics in the task lacks complexity; • students may need to identify patterns but are not pressed for generalizations; • students may use multiple strategies or representations but there is little emphasis on developing connections between them; • students may make conjectures but are asked to provide little or no mathematical evidence or explanations to support conclusions.
2	The potential of the task is limited to engaging students in using a procedure that is either specifically called for or its use is evident based on prior instruction, experience, or placement of the task. There is little ambiguity about what needs to be done and how to do it. The task does not require students to make connections to the concepts or meaning underlying the procedure being used. Focus of the task appears to be on producing correct answers rather than developing mathematical understanding (e.g., applying a specific problem solving strategy, practicing a computational algorithm).
1	The potential of the task is limited to engaging students in memorizing or reproducing facts, rules, formulae, or definitions. The task does not require students to make connections to the concepts or meaning that underlie the facts, rules, formulae, or definitions being memorized or reproduced. OR The task requires no mathematical activity.
N/A	**Reason:**

B. Implementation of the Task	
4	Students engage in using complex and non-algorithmic thinking or by

	exploring and understanding the nature of mathematical concepts, procedures, and/or relationships.*
3	Students engage in complex thinking or in creating meaning for mathematical procedures and concepts BUT the problems, concepts, or procedures do not require the extent of complex thinking as a "4"; OR The "potential of the task" was rated as a 4 but students only moderately engage with the high-level demands of the task .*
2	Students engage with the task at a procedural level. Students apply a demonstrated or prescribed procedure. Students may be required to show or state the steps of their procedure, but are not required to explain or support their ideas. Students focus on correctly executing a procedure to obtain a correct answer.*
1	Students engage with the task at a memorization level. Students are required to recall facts, formulas, or rules (e.g., students provide answers only). OR The task requires no mathematical activity.
N/A	**Reason:**

*See descriptors for "Potential of the Task" rubric on page 1 for examples.

C. Student Discussion Following the Task	
4	Students show/describe written work and provide complete and thorough explanations of **why** their strategy, idea, or procedure is valid. Students

	explain why their strategy works and/or is appropriate for the problem by making connections to the underlying mathematical ideas (e.g., "I divided because we needed equal groups"). OR Students show/discuss **more than one** strategy or representation* for solving the task, and provides explanations of **how/why** the different strategies/representations / mathematical ideas were used to solve the task and/or make connections between strategies / representations / mathematical ideas.
3	Students show/describe written work and attempt to provide explanations of why their strategy, idea, or procedure is valid. BUT the explanations are incomplete, incoherent, or lack precision (e.g., student responses often require extended press from the teacher). OR Students show/discuss **more than one** strategy or representation* for solving the task . Students may provide explanations of **how** the different strategies/representations were used to solve the task, but do not show connections nor explain why the strategy/representation was valid.
2	Students show/describe written work for solving the task (e.g., the steps for a multiplication problem, finding an average, or solving an equation; what they did first, second, etc) but do not explain **why** their strategy or procedure works and/or was appropriate for the problem; OR Students show/discuss only **one** strategy or representation* for solving the task.
1	Students provide brief or one-word answers (e.g., fill in blanks); OR Student's responses are non-mathematical.
N/A	**Reason:**

*Representations include numbers and/or symbols, diagrams/pictures, use of written/verbal language , graphs, tables/charts, concrete materials.]

D. Rigor of Expectations*:	
4	The majority of the teacher's observed expectations are for students to engage with the high-level demands of the task, such as **using complex thinking** and/or exploring and **understanding mathematical concepts,** procedures,

	and/or relationships.
3	At least some of the teacher's expectations are for students to engage in complex thinking or in understanding important mathematics. However, the teacher's expectations do not warrant a "4" because: • the expectations are appropriate for a task that lacks the complexity to be a "4"; • the expectations do not reflect the potential of the task to elicit complex thinking (e.g., identifying patterns but not forming generalizations; using multiple strategies or representations without developing connections between them; providing shallow evidence or explanations to support conclusions). • the teacher expects complex thinking, but the expectations do not reflect the mathematical potential of the task.
2	The teacher's expectations focus on skills that are germane to student learning, but these are not complex thinking skills (e.g., expecting use of a specific problem solving strategy, expecting short answers based on memorized facts, rules or formulas; expecting accuracy or correct application of procedures rather than on understanding mathematical concepts).
1	The teacher's expectations do not focus on substantive mathematical content (e.g., activities or classroom procedures such as following directions, producing neat work, or following rules for cooperative learning).
N/A	**Reason:**

*Rate this dimension based on the teacher's verbal directions, the task prompt, rubrics or criteria charts, modeling, etc.

Clear Expectations/Self- Management of Learning Observation Rubrics, 2003

Rate these dimensions holistically (not by individual student response)

I. Discussion (Lesson Task)

A. Clarity and Detail of expectations	
4	The expectations are very clear and explicit regarding the quality of work expected. The criteria for quality work are appropriately detailed.
3	The expectations are clear regarding the quality of work expected. However, there is no elaboration of what level of quality is expected for each criterion.
2	The expectations for the quality of student's work are broadly stated and un elaborated.
1	The teacher's expectations for the quality of student's work are unclear and/or unelaborated. OR the expectations for quality work are not shared with students.
N/A	Reason:

B. Access to expectations	
4	Criteria for the quality of work expected and how work will work will be scored is readily accessible to ALL students. There is a public record of these criteria.
3	Criteria for quality of work expected have been explicated to ALL students. However, there is no public record of these criteria.
2	Criteria for quality of work expected have been explicated to SOME students. There is no public record of these criteria.
1	The expectations for quality work are not shared with students.
N/A	Reason:

Rate these dimensions for each student interview

C. Understanding of expectations (Student Interview: Grade 1-2 only)	
4	Student clearly explains directions and expectations of quality for the task with details or examples. • Student explains what high, middle, and low-level performance looks like.
3	Student explains directions and expectations of quality for the task without much detail. • Student names a list of expectations.
2	Student vaguely explains directions and quality of expectations for the task. • Student just explains directions.
1	Student knows neither directions nor quality of expectations for the task
N/A	Reason:

Student A ____ Student B ____ Student C ____ Student D ____

II. Past Tasks (Student interview: all grades)

Rate these dimensions for each student interview

D. Judging work based on expectations	
4	Student clearly judges his/her own work based on the specific examples in the work. • Student demonstrates application of expectations to his/her own work (compares expectations to his/ her work) in detail. • Student translates general expectations to the task specifically.
3	Student judges his/her own work based on criteria in general terms. • Student attempts to apply expectations to his/her own work but general comparisons. • Students says, "I included this expectation."
2	Student vaguely judges his/her own work based on general terms. • Student points to expectations (e.g. scoring guide) but is unable to compare expectations to his/her work.
1	Student does not use the criteria to judge his own work
N/A	Reason:

Student A ____ Student B ____ Student C ____ Student D ____

E. Revising work based on expectations	
4	Student clearly explains his/her revision based on expectations with specific examples. • Student explains why s/he revised the work based on expectations and shows previous drafts and points to specific examples of revisions.
3	Student explains his/her revision based on expectations in general terms. • Student shows revisions and explains the reason in general terms based on expectations.
2	Student vaguely explains his/her revision without expectations. • Student shows revisions but doesn't explain the reasons based on expectations (e.g., "I did it to get a better grade or because the teacher told me to do so.")
1	Student is unable to explain his/her revisions or did not have the opportunity to revise his/her work.
N/A	Reason:

Student A ____ Student B ____ Student C ____ Student D ____

Academic Rigor - Math

F. Rigor of Expectations:	
4	The majority of the expectations described by the student are to engage with the high-level demands of the task, such as using complex thinking and/or exploring and understanding mathematical concepts, procedures, and/or relationships.
3	At least some of the expectations described by the student are to engage in complex thinking or in understanding important mathematics. However, the expectations do not warrant a "4" because: • the expectations are appropriate for a task that lacks the complexity to be a "4"; • the expectations do not reflect the potential of the task to elicit complex thinking (e.g., identifying patterns but not forming generalizations; using multiple strategies or representations without developing connections between them; providing shallow evidence or explanations to support conclusions). the teacher expects complex thinking, but the expectations do not reflect the mathematical potential of the task.
2	The expectations focus on skills that are germane to student learning, but these are not complex thinking skills (e.g., expecting use of a specific problem solving strategy, expecting short answers based on memorized facts, rules or formulas; expecting accuracy or correct application of procedures rather than on understanding mathematical concepts).
1	The expectations do not focus on substantive mathematical content (e.g., activities or classroom procedures such as following directions, producing neat work, or following rules for cooperative learning).
N/A	**Reason:**

Current Lesson Task (Grade 1-2 only): Student A ____Student B ____ Student C ____ Student D ____

Past Task: Student A ____ Student B ____ Student C ____
Student D ____

Observation Checklists, 2003

***Accountable Talk Function Checklist, 2003*: Check all that apply and script relevant contributions.**

Most of these moves will be made by the teacher, but in some cases, students might make them. In recording the actual moves, note T for Teacher move, S for Student move.

(script here)

1. Linking contributions

☐ Getting students to relate to one another's ideas
- "Jay just said…and Susan, you're saying that…"
- "Who wants to add on to what Ana just said?"
- "Who agrees and who disagrees with what Ana just said?"
- "How does what you're saying relate to what Juan just said?"
- "I agree with Sue, but I disagree with you, because…"
- S- "I agree with Fulano because…"

2. Accountability to knowledge

☐ Pressing for accuracy
- "Where could we find more information about that?"
- "Are we sure about that? How can we know for sure?"
- "Where do you see that in the text?"
- "What evidence is there?"
- T revoices S contribution and checks for accuracy

☐ Building on prior knowledge / recalling prior knowledge
- T or S links present work to past work
- "How does this connect with what we did last week?"
- "Do you remember when we read another book by this author?""

3. Accountability to rigorous thinking

☐ Pressing for reasoning

"What made you say that?"
"Why do you think that?"
"Can you explain that?"
"Why do you disagree?
"Say more about that."
"Let's let Fulano think."

Mathematics Observation Checklist, 2003

Check each box that applies:

A ↑ Lesson Activity provides **opportunities** for students **to engage** with the **high-level demands** of the task:	**B** ↓ During the Lesson Activity, the high-level **demands of the task** are **removed or reduced**:
▫ Students use multiple strategies and representations. ▫ Students communicate mathematically with peers. ▫ Teacher provides scaffolding that supports students to engage with the high-level demands of the task while maintaining the challenge of the task. ▫ Teacher provides sufficient time to grapple with the demanding aspects of the task and for expanded thinking and reasoning. ▫ Teacher holds students accountable for high-level products and processes. ▫ Teacher provides consistent presses for explanation and meaning. ▫ Teacher provides students with sufficient modeling of high-level performance on the task. ▫ Teacher provides encouragement for students to make conceptual connections. ▫ Other:	▫ Students are not pressed or held accountable for high-level products and processes or for explanations and meaning. ▫ The task is not complex enough to sustain student engagement in high-level thinking. The scaffolding is too directive and serves to remove or reduce the challenging aspects of the task: ▫ Teacher provides a set procedure for solving the task ▫ The focus shifts to procedural aspects of the task or on correctness of the answer rather than on meaning and understanding. ▫ Feedback, modeling, or examples are too directive or did not leave any complex thinking for the student. Students are not provided with enough scaffolding to make or sustain progress on the task: ▫ Students are not given enough time to deeply engage with the task or to complete the task to the extent that was expected. ▫ Students do not have the prior knowledge necessary to engage with the task at a high level. ▫ Students do not have access to resources necessary to engage with the task at a high level. ▫ Other:

C The **Discussion** provides opportunities for students to engage with the high-level demands of the task:
▫ Students use multiple strategies and make explicit connections or comparisons between these strategies, or explain why they choose one strategy over another. ▫ Students use or discuss multiple representations and make connections between different representations or between the representation and their strategy, underlying mathematical ideas, and/or the context of the problem ▫ Students identify patterns or make conjectures, predictions, or estimates that are well grounded in underlying mathematical concepts or evidence. ▫ Students generate evidence to test their conjectures. Students use this evidence to generalize mathematical relationships, properties, formulas, or procedures. ▫ Students (rather than the teacher) determine the validity of answers, strategies or ideas. ▫ Other:

Clear Expectations/ Self-Management of Learning Observation Checklist, 2003

Clear Expectations / Self-Management of Learning (CE/SML)

Means of communicating expectations <u>during the lesson</u>

Check all below that were used to communicate expectations during the lesson.

☐ Criteria chart
☐ Process chart
☐ Rubric
☐ Model of student performance that meets standard
☐ Model of intermediate expectation
☐ Counter-model of unacceptable performance
☐ Template- outlines all the steps and information necessary to complete the task
☐ Oral explanation of expectations
☐ Other: _____

Means of communicating expectations <u>during the student interviews</u>

Check all below that were used to communicate expectations during student interviews. Ask students about these means of communicating expectation with students during interviews. Photograph relevant charts, handouts, etc.

☐ Criteria chart
☐ Process chart
☐ Rubric
☐ Model of student performance that meets standard
☐ Model of intermediate expectation
☐ Counter-model of unacceptable performance
☐ Template- outlines all the steps and information necessary to complete the task
☐ Oral explanation of expectations
☐ Other: _____

Mathematics Assignment Rubrics, 2003
Dimension 1

Academic Rigor: Potential of the Task

Rubric 1: Potential of the Task	
4	The task has the potential to engage students in "doing mathematics" or "procedures with connections" : • using complex and non-algorithmic thinking (i.e., there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instructions, or a worked-out example); • exploring and understanding the nature of mathematical concepts, procedures, and/or relationships. The task may require students to: • solve a genuine, challenging problem; • develop an understanding for why formulas or procedures work; • apply a broad general procedure that remains closely connected to mathematical concepts; • identify patterns and form generalizations based on these patterns; • make conjectures and support conclusions with mathematical evidence; • make connections between representations, strategies, or mathematical concepts and procedures.
3	The task has the potential to engage students in complex thinking or in creating meaning for mathematical concepts, procedures, and/or relationships. However, the task does not warrant a "4" because: • students engage in problem solving, but the mathematics in the task lacks complexity; • students engage in cognitively not challenging task; the task is easy to solve • students may need to identify patterns but are not pressed for generalizations; • students may use multiple strategies or representations but there is little emphasis on developing connections between them; • students may make conjectures but are asked to provide little or no mathematical evidence or explanations to support conclusions.
2	The potential of the task is limited to engaging students in using a procedure that is either specifically called for or its use is evident based on prior instruction, experience, or placement of the task. There is little ambiguity about what needs to be done and how to do it. The task does not require student to engage in cognitively challenging work; the task is easy to solve. The task does not require students to make connections to the concepts or meaning underlying the procedure being used. Focus of the task appears to be on producing correct answers rather than developing mathematical understanding (e.g., applying a specific problem solving strategy, practicing a computational algorithm).
1	The potential of the task is limited to engaging students in memorizing or reproducing facts, rules, formulae, or definitions. The task does not require students to make connections to the concepts or meaning that underlie the facts, rules, formulae, or definitions being memorized or reproduced. OR The task requires no mathematical activity.

*Representations include numbers and/or symbols, diagrams/pictures, use of written/verbal language , graphs, tables/charts, concrete materials.]

Dimension 2

Academic Rigor: Implementation

Rubric 2: Implementation of the Task	
4	Student-work indicates use of complex and non-algorithmic thinking, problem solving, or exploring and understanding the nature of mathematical concepts, procedures, and/or relationships.*
3	Students engage in problem-solving or in creating meaning for mathematical procedures and concepts BUT the problems, concepts, or procedures do not require the extent of complex thinking as a "4"; OR The "potential of the task" on page 1 was rated as a 4 but Ss only moderately engage with the high-level demands of the task.*
2	Students engage with the task at a procedural level. Students apply a demonstrated or prescribed procedure. Students may be required to show or state the steps of their procedure, but are not required to explain or support their ideas. Students focus on correctly executing a procedure to obtain a correct answer.*
1	Students engage with the task at a memorization level. Students are required to recall facts, formulas, or rules (e.g., students provide answers only). OR Students do not engage in mathematical activity.

Dimension 3
Academic Rigor: Discussion

Rubric 3: Student Discussion Following Task	
4	Students show written work and provide complete and thorough explanations of why their strategy, idea, or procedure is valid. Students explain why their strategy works and/or is appropriate for the problem by making connections to the underlying mathematical ideas (e.g., "I divided because we needed equal groups"). OR Student work displays use of more than one strategy or representation* for solving the task, and provides a written explanation of how the different strategies/representations were used to solve the task.
3	Students show written work and provide explanations BUT the explanations are incomplete or are procedural in nature. Students explain the steps of their work (e.g., what they did first, second, etc.) but do not explain **why** their strategy or procedure works and/or was appropriate for the problem; OR Student work displays use of more than one strategy or representation* for solving the task.
2	Students show written work for solving the task (e.g., the steps for a multiplication problem, finding an average, or solving an equation) with no written explanation; OR Student work displays use of only one strategy or representation* for solving the task.
1	Students provide brief or one-word answers (e.g., fill in blanks); OR Student's responses are non-mathematical.

Academic Rigor: Expectations

Rubric 4: Academic Rigor in Teacher's Expectations*	
4	The majority of the teacher's expectations are for students to: • use complex and non-algorithmic thinking (i.e., there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instructions, or a worked-out example); • explore and understand the nature of mathematical concepts, procedures, and/or relationships. [The expectations for mathematical content are stated explicitly in one of the sources indicated by the * below.] For example, the teacher may expect students to: • solve a genuine, challenging problem; • develop an understanding for why formulas or procedures work; • identify patterns and form generalizations based on these patterns; • make conjectures and support conclusions with mathematical evidence; • make connections between representations, strategies, or mathematical concepts and procedures.
3	At least some of the teacher's expectations are for students to engage in complex thinking or in understanding important mathematics. However, the teacher's expectations do not warrant a "4" because: • the expectations are appropriate for a task that lacks the complexity to be a "4"; • the expectations do not reflect the potential of the task to elicit complex thinking (e.g., identifying patterns but not forming generalizations; using multiple strategies or representations without developing connections between them; providing shallow evidence or explanations to support conclusions). • the teacher expects complex thinking, but the expectations do not reflect the mathematical potential of the task.
2	The teacher's expectations focus on skills that are germane to student learning, but these are not complex thinking skills (e.g., expecting use of a specific problem solving strategy, expecting short answers based on memorized facts, rules or formulas; expecting accuracy or correct application of procedures rather than on understanding mathematical concepts).
1	The teacher's expectations do not focus on substantive mathematical content. The teacher's focus may be solely on activities or classroom procedures (e.g., following directions, producing neat work, or following norms for cooperative learning).

*Rate this dimension based on Coversheet Q 4 and the attached rubric.

Dimension 1
Clear Expectations: Clarity and Detail of Expectations

Rubric 1: Clarity and Detail of Expectations	
4	The expectations for the quality of students' work are very clear and elaborated. Each dimension or criterion for the quality of students' work is clearly articulated. Additionally, varying degrees of success are clearly differentiated.
3	The expectations for the quality of students' work are clear and somewhat elaborated. Levels of quality may be vaguely differentiated for each criterion (i.e., little information is provided for what distinguishes high, medium and low performance.)
2	The expectations for the quality of student's work are broadly stated and unelaborated.
1	The teacher's expectations for the quality of student's work are unclear OR the expectations for quality work are not shared with students.

Dimension 2

Clear Expectations: Communication of Expectations

Rubric 2: Communications of Expectations	
4	Teacher discusses the expectations or criteria for student work (e.g., scoring guide, rubric, etc.) with students in advance of their completing the assignment and models high-quality work.
3	Teacher discusses the expectations or criteria for student work (e.g., scoring guide, rubric, etc.) with students in advance of their completing the assignment.
2	Teacher provides a copy of the criteria for assessing student work (e.g., scoring guide, rubric, etc.) to students in advance of their completing the assignment.
1	Teacher does not share the criteria for assessing students' work (e.g., scoring guide, rubric, etc.) with the students in advance of their completing the assignment. (e.g., Teacher may provide a copy of the scoring rubric to students when giving them their final grade.
N/A	Reason: