Using the Instructional Quality Assessment (IQA) Toolkit to Assess Academic Rigor in Mathematics Lessons and Assignments

Melissa Boston and Mikyung Kim Wolf,

University of Pittsburgh, Learning Research and Development Center

Paper presented at the Annual Meeting of the American Educational Research Association Meeting, San Diego, April 2004

Contact Information:

Melissa Boston

University of Pittsburgh, Learning Research and Development Center

Office #727

3939 O'Hara Blvd.

Pittsburgh, PA 15260

(412) 624-2150

melissab@pitt.edu

Introduction

This paper extends the discussion of the Instructional Quality Assessment (IQA) toolkit presented by the earlier papers in this symposium to focus specifically on the Academic Rigor in Mathematics (AR-Math) Lesson Observation and Assignments rubrics. The IQA endeavored to design rubrics that would measure the quality of instruction and learning in school mathematics programs. Such assessments could inform mathematics program directors at the school level, indicate areas in need of professional development for teaching staff, and serve as professional development tools for teachers. The purpose of this paper is to describe the development of the Academic Rigor rubrics in Mathematics for lesson observations and for assignments, to share the findings from a small scale 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 paper is intended to answer the question, "Is the IQA toolkit a reasonable means of assessing the academic rigor of school mathematics programs?" Specifically, the discussion focuses on (1) the rater reliability of the AR-Math rubrics, (2) the relationships among the AR-Math rubrics, and (3) the detective power of the IQA to distinguish the quality of mathematics instructional programs. This discussion begins by describing the theoretical basis of the AR-Math rubrics in the following section.

Theoretical Basis of the IQA Academic Rigor in Mathematics Rubrics

The Principle of Learning entitled *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 (IFL, 2002). In essence, academic rigor in mathematics refers to 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 (Hiebert & Carpenter, 1992; Putnam, et al., 1990). 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 & Carpenter (1992) contend 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, & 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 (Voigt, 1994).

Consistent with the theories noted above, 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.

Opportunities for students to learn mathematics with understanding can be determined by assessing the cognitive demands of the instructional tasks that students engage with during mathematics instruction. A ***mathematical task*** is defined as a set of problems or a single complex problem that focuses students' attention on a particular mathematical idea (Stein, Grover, & Henningsen, 1996). The construct of task also encompasses the intellectual and physical products that are expected of students, the operations that students are to use to obtain the desired products, and the resources that are available to assist students in completing these products (Doyle, 1983). ***Cognitive demands*** describe the types of thinking that students engage in while solving a task. The types of thinking that generate learning mathematics with understanding are referred to as ***high-level*** cognitive demands. The following section will describe the ways in which mathematical tasks influence students' opportunities to learn mathematics with understanding, and thus provide a valid indicator of academic rigor in mathematics.

<u>The Influence of Tasks on Students' Learning.</u>

On a practical level, tasks influence student learning because working on mathematical tasks constitutes what students *do* during the majority of their time in mathematics class. Students in all seven countries analyzed in the TIMSS Video Study (USDE, 2003) (i.e., Australia, Czech Republic, Hong Kong, Japan, Netherlands, Switzerland, and the United States) spent over 80% of their time in mathematics class working on mathematical tasks. 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. Tasks thus influence learning by defining the mathematical content that students have an opportunity to learn. 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, Carpenter, Fennema, Fuson, Wearne, Murray, Olivier, & Human, 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 result in much different learning than a task that requires problem-solving, conjecturing, and reasoning. Mathematical tasks with high-level cognitive demands (hereafter referred to as ***high-level tasks***) 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, 1997). Hiebert and colleagues (Hiebert, et al., 1997; Hiebert and 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 and colleagues (Putnam, et al., 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 and colleagues (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 et al, 1996; Doyle, 1983). Tasks with low levels of cognitive demand (hereafter referred to as ***low-level tasks***) should not unambiguously be considered 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). Research to support this claim will be provided in the following section.

The Relationship Between High-Level Tasks and Student Learning

A growing body of research indicates that curricular materials specifically developed to contain high-level tasks (USDE, 1999) are successful in improving students' performance on

state and national tests of mathematical achievement (e.g., Fuson, Carroll, & Druek, 2000; Riordan and 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 and Senk, 2000; Reys, et. al, 2003), and in improving students abilities to reason, communicate, problem-solve and make mathematical connections (e.g., Ridgeway, Zawojewski, Hoover, & Lambdin, 2003; Schoenfeld, 2002). Exposing students to tasks that embody high-level cognitive demands appears to have a positive effect on students' development of mathematical understanding.

In their analysis of mathematics lessons from teachers participating in the QUASAR Project[1], Stein and Lane (1996) investigated the degree to which variations in the implementation of reform-oriented features of mathematics instruction could be linked to variations in students' learning. Stein and Lane (1996) found that instruction characterized by the use of tasks featuring high-level cognitive demands, multiple solution strategies, multiple representations, and opportunities for mathematical discourse generated greater student learning gains than classrooms where the mathematical tasks did not exhibit these characteristics. This result suggests that the potential cognitive demands of a mathematical task are an important indicator of students' opportunities to learn mathematics with understanding.

[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).

Task Implementation

Cognitive demands can and often do change over the course of an instructional episode (Henningsen & Stein, 1996). Teachers and students accustomed to traditional, directive styles of teaching and routinized, procedural tasks experience conflict and discomfort with the ambiguity and struggle that often accompany high-level tasks (Smith, 1995; Clarke, 1997). In response to ambiguity or uncertainty on how to proceed, students may disengage with the task or press the teacher for step-by-step instructions (Romanagno, 1994; Henningsen & Stein, 1996) -- thereby reducing the cognitive demands of the task. Teachers in the QUASAR Project varied in their ability to maintain the cognitive demands of high-level tasks (Stein & Lane, 1996), leading QUASAR researchers to identify factors that both supported and inhibited the maintenance of high-level cognitive demands throughout an instructional episode (Henningsen & Stein, 1996). Further evidence that maintaining the complexity of high-level tasks is not a trivial endeavor is provided by the finding that, of the 17% of tasks identify as high-level by TIMSS researchers, less that 1% of those tasks were enacted in ways that provided evidence of mathematical connections, reasoning, or problem-solving (USDE-NCES, 2003). Conversely, findings from the TIMSS 1999 Video Study also provides examples of lessons in which the cognitive demands of a task increased during implementation. Tasks that were considered as having low-level cognitive demands (i.e., tasks that required students to state concepts or procedures from memory) were enacted during the implementation of the lesson in ways that required students to make connections between mathematical concepts (USDE-NCES, 2003).

Research shows that the greatest student learning gains occur in classrooms in which high-level demands are consistently maintained. Stein and Lane (1996) concluded that, in addition to being exposed to high-level tasks, students also needed to engage with the high-level aspects of the tasks throughout the instructional episode. These results appear consistent with

findings from the TIMSS 1999 Video Study (USDE-NCES, 2000), in which higher performing countries were found to implement high-level tasks in ways that maintained the high-level cognitive demands. Assessing the implementation of the task is thus essential to ascertaining academic rigor in mathematics instruction.

One of the features 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 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. Hiebert and colleagues (1997) describe these types of opportunities for students to reflect and communicate about their mathematical work as essential for learning mathematics with understanding.

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; Clake, 1997; Romanagno, 1994). Thus, a teacher's expectations for students' learning can influence what happens during an instructional episode and throughout students' experiences in that teachers' classroom. Henningsen and Stein (1996) 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 not likely to spontaneously do more than 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 depending on what they perceive is required by the task and by the teacher (Doyle, 1983). Teacher's expectations determine what students will be held

accountable for, and this accountability frames the work that students will choose to engage in during instruction. In this way, a teacher's expectations for students' work influences the academic rigor of mathematics instruction.

This section has identified four important aspects of students' opportunities to learn mathematics with understanding: high-level tasks, implementation of high-level tasks, mathematical discussions following students' work on tasks, and teacher's expectations. The dimensions of the IQA AR-Math rubrics were designed to reflect these aspects of academic rigor in mathematics instruction. The development of the AR-Math rubrics will be described in the following section.

Development of the IQA Academic Rigor in Mathematics Rubrics

The IQA AR-Math rubrics for both Lesson Observations and Assignments were created to consist of four dimensions critical to assessing students' opportunities to learn mathematics with understanding: potential of the task, implementation of the task, student discussion or students' written responses; and teachers' expectations. Based on research stemming from the QUASAR project (i.e., Stein, Grover, & Henningsen, 1997; Henningsen & Stein, 1996), 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 teachers' expectations for high-level cognitive processes (teacher's expectations), and in the cognitive processes evident in the discussion or in students' written responses to the assignment. In each dimension, descriptors for score levels 3 and 4 are consistent with characteristics of high-level cognitive demands. These levels differ with regard to (1) the complexity of the task or of the mathematics in the task and (2) the explicitness of the mathematical connections or reasoning. Score levels 1 and 2 reflect low-level cognitive demands

as described by Stein, Grover, and Hennignsen (1997); with level 2 resonant of "procedures without connections" and level 1 of "memorization" or "no mathematical activity." Hence, an important demarcation line exists between the score levels of 2 and 3 that separates high- and low- level cognitive demands in each dimension of the AR-Math rubrics.

Each dimension is described more specifically below:

- 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, & Henningsen (1997). 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).
- 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* elicit from students, the "Implementation" dimension assess 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 level at which the task was implemented during the lesson, including both student work time and any whole or small group discussion. Task demands can be altered during instruction, and certain instructional factors serve to maintain or to reduce students' opportunities to engage in high-level cognitive processes as they engage with mathematical tasks (Henningsen & Stein, 1996). A Mathematics Lesson

Checklist based on the factors identified by Henningsen & Stein is also provided to guide the scoring of the Implementation dimension.

- AR3: Rigor in Student Discussion or Responses. The dimension of Rigor in the Discussion Following the Task (hereafter referred to as Student Discussion) assess 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.
- 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. The Clear Expectations rubrics (CE) assess the clarity of the teacher's expectations for students; this dimension assesses the level of cognitive demand in the teacher's expectations for students' work on the task.

The previous sections have presented the theoretical underpinnings that form the basis of the IQA AR-Math rubrics and described the development of the dimensions in the rubrics. The purpose of providing this information is to make the claim that the AR-Math rubrics assess important aspects of students' opportunities to learn mathematics with understanding. The remainder of the paper will focus on a small scale pilot study testing the usability and the technical quality of the IQA toolkit. As noted in the paper by Matsumura, Wolf, and Crosson (2004) presented earlier in this symposium, the results and conclusions from this pilot study must be interpreted cautiously due to the small sample size. The results have, however, provided the IQA development team with valuable information to guide future development of the toolkit.

The IQA Spring 2003 Pilot Study

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." These schools served a diverse population of students (26% African American, 6% Asian, 47% Latino, 15% white, 6% other), 20% of whom were English language learners. Teachers who participated in the study had been teaching for an average of 14 years, and had been at their school an average of 4 years (Matsumura, et al., 2004). Fifteen teachers participated in the mathematics

portion of the study, and 14 of these teachers turned in assignments with samples of student work.

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 IFL in which they had been involved in consistent and sustained efforts to implement the Principles of Learning into their schools and classrooms, while 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 paper) that contained a predominance of mathematical tasks that involved thinking, reasoning, and sense-making (i.e., high-level tasks). The elementary mathematics curriculum in District D had a more traditional focus of improving students' accurate and efficient execution 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.

Data

Data for each teacher consist of one lesson observation and four mathematics assignments. 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. Teachers also submitted six samples of student work for each task--two

samples of work they considered to be of high, medium and low quality respectively. The lesson observations were coded using the Accountable Talk (AT), Clear Expectations (CE), and Academic Rigor (AR) rubrics for lesson observations in mathematics and received a score of 1-4 in each dimension. The lesson observation also included a checklist to inform raters' scores for the *implementation* and *discussion* dimensions. The assignments were coded using the CE and AR rubrics for assignments in mathematics.

The lesson observations and assignments were rated by graduate students who were recruited to participate in the data collection and were not part of the team who developed the rubrics ($N = 2$). As described in the paper by Matsumura, Wolf, and Crosson (2004), "naïve" raters were used in order to assess the quality of our rater training program and rubrics (as evidenced by the degree to which people who were external to the project could agree on the different ratings). The assignments were randomly ordered for scoring and were rated independently by each of the naïve raters ($N = 54$ assignments).

Analysis

First, several measures were used to determine inter-rater reliability. Percent of exact agreement between the two raters was first computed. Cohen's kappa coefficients were calculated to investigate the level of agreement between two raters on each dimension when controlled for chance agreement. Correlations also were 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, generalizability studies were 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. Decision studies were conducted to explore options for future research designs. 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 and Discussion

This section presents and discusses the results from the analyses noted in the previous section. Reliability results are presented first, followed by descriptive statistics, generalizability studies, and inter-correlations.

Reliability.

The two raters scored the lesson observations and assignments independently, then engaged in a debriefing session in which a consensus was reached on a final score for each dimension. Reliability tests were conducted to compare the agreement of the two raters' initial 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 (47.6%), though 1-point agreement was excellent (95.2%). For individual dimensions within the AR-Math rubric, reliability ranged from poor to fair, with percentages of exact agreement between 33% for the *Potential of the Task* dimension (AR1) and 66.7% for the *Discussion* dimension (AR3). The correlation between raters was also poor ($\underline{r}$ = .34 to $\underline{r}$ = .41) for each dimension except *Discussion* ($\underline{r}$ = .71). 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 dimensions (ranging from 60.0% to 67.3%). The reliability results for Lesson Observations and Assignment ratings are presented in Table 1.

Table 1.
Inter-rater reliability for AR-Math Rubrics for Lesson Observation and Assignment scores.

AR-Math Dimensions	Lesson Observation (N = 15 teachers)			Assignments (N = 14 teachers)		
	% of exact agreement	Kappa	Spearman *r*	% of exact agreement	Kappa	Spearman *r*
AR1: Potential	33.3	-	.38	65.5	.51	.73
AR2: Implementation	46.7	.23	.41	60.0	.43	.72
AR3: Discussion*	66.7	.50	.71	67.3	.53	.74
AR4: Expectations	53.3	.33	.34	62.7	.43	.68

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

A great deal of disagreement about the potential of the task dimension surfaced during the rater debriefing for the mathematics lesson observations. Hence, low reliability in that dimension 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 lead to greater reliability.

Reliability was also analyzed by collapsing each dimension to a 3-point scale. Rater agreement after regrouping the 4-pont 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 dimensions (see Table 3).

Table 2.
Inter-rater Reliability of Lesson Scores after Regrouping Score Scales for Overall Math Rubrics (*N* = 15 teachers)

	% of exact agreement	Kappa
4 point scale (1-4)	47.6	.29
3 point scale (1, 2, 3/4)	64.6	.36
3 point scale (1, 2 /3, 4)	63.0	.36
3 point scale (1/2, 3, 4)	56.1	.34

Table 3.
Exact Agreement of Lesson Scores after Regrouping AR-Math Rubrics (1, 2, 3 /4) (*N* = 15 teachers)

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

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. One source of confusion appeared to lie between the score levels of 3 and 4, as evidenced by the improved reliability when levels 3 and 4 were combined. In each dimension, 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 dimension. The improved reliability results for combining levels 2 and 3 are cause for even greater concern. 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. This finding also has specific implications for rater training. Because the main difference between levels 2 and 3 is whether the mathematical procedure is connected to meaning and understanding and whether students are free to choose their own strategy (or representation) or one is provided for them, raters once again need additional 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.

Descriptive Statistics for Lesson Observations and Assignment Scores.

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 moderately higher in District C than in District D. These results suggest 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 were 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 (USDE-NCES, 2003).

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 histograms in Appendix A for the lesson observation scores in each dimension provide another portrayal of the differences in students' opportunities to learn in each of the districts.

Table 4.
Descriptive Statistics of Scores on AR-Math Lesson Observation Rubrics by Districts (*N* = 15 lessons).

AR Dimension	District C Mean (SD)	District D Mean (SD)	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 **N = 10

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. A 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.
Descriptive Statistics of Scores on AR-Math Assignment Rubrics by Districts (N = 54 assignments, 14 teachers).

AR Rubrics	District C Mean (SD)	District D Mean (SD)	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*

*$p < .05$

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 three Principles of Learning (AT, CE, and AR), 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.

Generalizability.

Generalizability and decision studies were conducted to determine how many raters and assignments might be necessary to obtain a stable estimate of the quality of classroom practice. Results indicated that our design based on two raters and four teacher assignments yielded an excellent generalizability 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 teacher and the assignment type interaction, suggesting that the teachers submitted different types of assignments. However, compared to the results from the analysis of reading comprehension assignments reported in earlier paper by

Matsumura et al. (2004), the assignment types appeared to be consistent per teacher in mathematics.

Table 6.
Estimates of Variance Components (N = 14 teachers) for the Mathematics Assignments

Source of Variation	Estimated Variance Component [a]	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

[a]. Negative variance component was set to zero.

Table 7 presents the results of decision studies indicating that varying the number of assignments (from 3 to 6) and the number of raters (from 1 to 2) would also generate stable estimates of the quality of classroom practice. These results lend support to the contention that students' assignments can be used as indicators of classroom practice (Clare & Aschbacher, 2001). The results can also inform the design of other research intending to utilize student work to assess students' opportunities to learn mathematics with understanding. Note, however, that these findings differed substantially from the results of the generalizability and decision studies for reading comprehension assignments discussed by Matsumura, Wolf, and Crosson (2004). Perhaps this is due to the consistent nature of the tasks within the mathematics curricular

materials in each district, or due to the influence of teacher's expectations on the types of tasks they would select and provide to their students as assignments.

Table 7.
Estimated G-Coefficients for Mathematics Assignments. (*N* = 14 teachers)

Number of Assignments	Number of Raters	Estimated G-Coefficient
3	1	0.84
4	1	0.87
5	1	0.89
3	2	0.89
4	2	0.91
5	2	0.93
6	2	0.94

Correlation of Rubric Dimensions.

Results indicate that all four AR-Math dimensions were significantly correlated within Lesson Observation rubrics and within Assignment rubrics. These results are provided in Tables 8 and 9. Particularly, the *Potential of the Task* and the *Teacher's Expectations* are highly correlated for the mathematics rubrics, as well as for the Reading Comprehension rubrics (Matsumura, et al., 2004). Further investigation would be needed to ascertain whether this correlation indicates a redundancy in rubric dimensions or a desired outcome of the rubrics. In other words, answers to the question of whether certain dimension can be eliminated might differ based on statistical relevance vs. practical relevance in informing mathematics instruction at the school or teacher level. What would be the instructional implications if teacher's expectations were not consistent with the potential of the task? 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?

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.

Conclusions from the pilot study and potential implications for professional development for mathematics teachers using the IQA toolkit will be provided in the closing section.

Table 8.
Inter-correlation within Lesson Observation Scores on AR: Math Rubrics.

	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 9.
Inter-correlation within Assignment Scores on AR: Math Rubrics.

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

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

Conclusions

Is the IQA Toolkit an effective tool for assessing academic rigor in school mathematics programs?

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 were frequently 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. Interpreting this result with the high correlation between the potential of the assignment tasks and teacher's expectations seems to indicate that individual teachers tend to give assignments of consistent score levels that reflect the level of their expectations for students' learning. Hence, raising teacher's expectations can in turn generate increases in students' opportunities to engage with high-level tasks – thereby increasing students' opportunities to learn mathematics with understanding.

Overall, we contend that the IQA Toolkit identifies important differences in students' opportunities to learn mathematics. Of course, we would have like better reliability, but we are optimistic that reliability will improve with increased rater training. Supplements to rater training in mathematics will specifically focus on the P*otential of the Task* dimension, on identifying evidence of mathematical connections and reasoning, and on general differences between high and low-level cognitive demands in mathematics. The paper will close by offering a suggestion for future use of the IQA toolkit: the professional development of mathematics teachers.

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

In choosing to focus on the mathematical tasks that teachers engage students with during mathematics instruction, the IQA rubrics draw on research and theories ascertaining that the most important first step in providing students with enriched opportunities to learn mathematics is to enrich the tasks that students engage with during mathematics instruction (Stein & Lane, 1996; Doyle, 1988). 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, worthwhile, or appropriate 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 students 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

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.

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., and 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.

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

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: Erlbaum.

National Council of Teachers of Mathematics (NCTM) (2000). Principles and Standards for School Mathematics. Reston, Va.: NCTM.

National Council of Teachers of Mathematics. (1989). Curriculum and Evaluation Standards for School Mathematics. Reston, Va.: NCTM.

National Council of Teachers of Mathematics. (1991). Professional Standards for Teaching Mathematics. Reston, Va.: NCTM.

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, D.C.: AERA.

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

Ridgeway, J.E, Zawojewski, J.S., Hoover, M.N., & Lambdin, D.V. (2003). Student attainment I 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. 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.

Smith, M.S. (1995). A Road to Change. (Doctoral dissertation, University of Pittsburgh, 1995). UMI Dissertation Services, 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.

USDE (1999). Exemplary and promising mathematics programs: Expert panel report. www.enc.org/professional/federalresources/exemplary/promising/

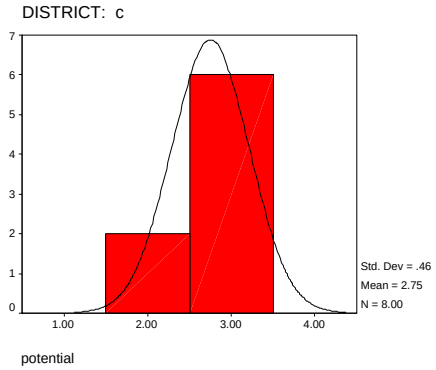
USDE, NCES (2003). Teaching mathematics in seven countries: Results from the TIMSS 1999 video study. NCES (2003-013). Washington, DC: Author.

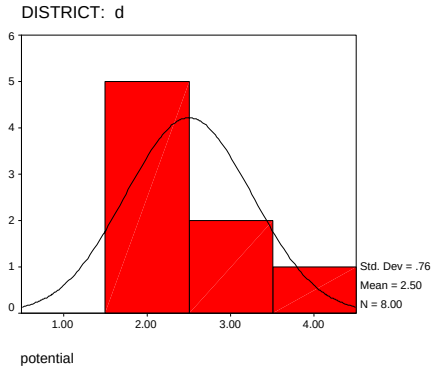
Voigt, J. (1994). Negotiation of mathematical meaning and learning mathematics. Educational Studies in Matheamtics, 26, 275-298.

Appendix A

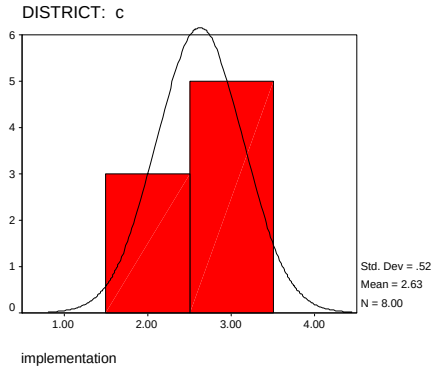
Histograms for AR-Math Lesson Observation Rubric Dimensions by District

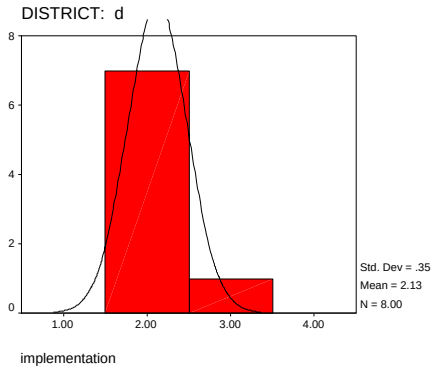
AR1: Potential



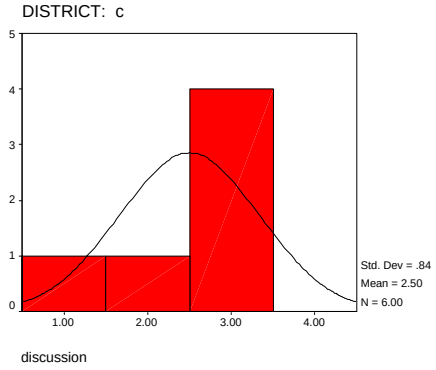


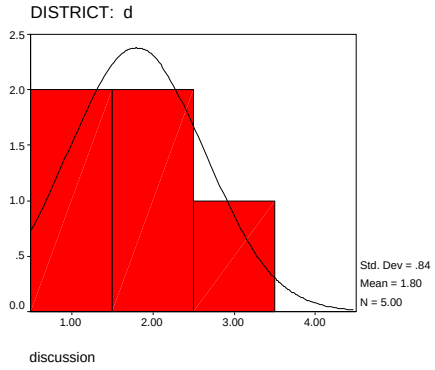
AR2: Implementation





AR3: Discussion





AR4: Expectations