

Piloting Math Benchmark Assessments

A Report for the CPS Office of Mathematics and Science
Prepared by the PRAIRIE Group, UIC College of Education

Lisa M. Raphael Bogaert, Carol R. Fendt, & Erica Vogele

June 15, 2006

For further information, contact Carol R. Fendt at crfendt@hotmail.com, 312-413-3367.

The conclusions drawn in this report reflect the viewpoints of the authors. While there are many potential viewpoints, these reflect a systematic analysis of data by external evaluators. The hope is that these findings can facilitate improvement of this and related programs through open discussion and consideration of data-driven understandings.

This report is based upon work supported by the Chicago Community Trust and the Chicago Public Schools.

Abstract

The purpose of this report is to describe the piloting of the Math Benchmark Assessments, a formative assessment system developed by the Center for the Study of Learning, Instruction, and Teacher Development (LITD) at the University of Illinois at Chicago along with the Office of Mathematics and Science (OMS) at the Chicago Public Schools (CPS). This joint venture began in 2004-05 when LITD in collaboration with the OMS developed the assessment system. (Information on the development of these assessments is available on the OMS website at <http://cmsi.cps.k12.il.us/ViewProgramDetails.aspx?pid=157>, *Review of Assessments Contained in CMSI-Supported Curricula*. The 2005-06 pilot testing of the Math Benchmark Assessments in CPS elementary schools occurred officially in CPS Instructional Areas 9 and 12. The data collection for the evaluation of this program began in the 2005-06 school year.

Introduction

As part of its effort to improve mathematics teaching and learning in the Chicago Public Schools (CPS), the Office of Math and Science (OMS) in collaboration with the Center for the Study of Learning, Instruction and Teacher Development (LITD) at the University of Illinois at Chicago (UIC) implemented a field test of a formative Benchmark Assessment program during the 2005-06 academic year. This field test consisted of three separate five-week rounds when teachers assessed their students and then received feedback on the results of these assessments. These three rounds took place across the school year in October, January, and May.

The purpose of the Math Benchmark Assessment program was to provide regular, instructionally relevant information about student performance in mathematics to teachers in a timely way. The ultimate goal was to promote discussion among teachers, students,

principals, and other staff members about student work in mathematics. The overarching purpose was to contribute to a coherent assessment system to meet the needs of educators and policymakers.

The Benchmark Assessments were intended and designed to be formative, low-stakes assessments; to be used solely for helping teachers gain a better idea of how their students understand key mathematical concepts. The results were to inform their planning of further learning activities in the classroom and beyond. It was also to serve as an intermediate link between the mathematics curriculum being implemented in schools and the state and district-wide goals for mathematics learning.

In order to determine how to best support the use of the math Benchmark Assessments, the OMS of CPS piloted the use of these assessment tools and tried two different strategies for how to support the process. In one geographic-based Instructional Area, the plan was for all teachers at participating schools to be directly instructed in the concept of Benchmark Assessments and the use of the materials provided. This was the “train-the-teacher” model. In a second Area, only school-based math specialists or lead teachers at participating schools were trained to use the assessments with students and then score their work. The plan was that these school leaders would then return to their schools and train the rest of the staff in how to do this work. This was the “train-the-trainer” model. Twenty-seven schools in CPS Elementary Instructional Areas 9 and 12 participated in this pilot test process in which they used the assessments in each classroom at each grade (3-8).

In the course of this pilot test for this program, the UIC LITD and CPS OMS created a variety of tools. The actual math assessments administered to students were in the form of twelve multiple choice questions two extended response questions, and one short answer question (added to the test after the first administration). Each grade had two forms of the test; however, the only test items different between tests were the extended response questions. For ease of use and to reinforce the idea that this was not a “high stakes” exam, these tests were to be given in a regular class period and teachers were given the option to give it whenever they wanted within a certain period of time (e.g. April 26-May 4). In addition to the test booklets to be used with the students, other necessary tools were created including:

- administration manuals for teachers using the assessments
- coordinator manuals and presentations
- answer keys for scoring
- scoring guides and rubrics
- multiple choice distractor analysis
- documents mapping items to IL Standards
- curricula for three half-day long professional development sessions for teachers and specialists/teacher-leaders

All of these tools were made available on the CMSI website shortly after they were developed.

Evaluation Methods

This report has been created by external evaluators (PRAIRIE Group of UIC) working both for the LITD whose work on this is partially funded by a grant from the Chicago Community Trust and for the CPS OMS which has hired these external evaluators to look at their overall efforts in improving math and science education in elementary schools.

LITD, OMS, and the external evaluators negotiated the questions guiding this evaluation at the start of the 2005-06 school year. These questions included the following:

1. Early Process of Enacting the Program

- a. How does the professional development take place given standards of good professional development?
- b. What are the messages delivered in terms of project goals in various forums of workshops, written tools, interactions among key program leaders and school participants?
- c. In what ways do school activities around Benchmark Assessments vary?
- d. What supports/barriers assist the use of the Benchmark Assessments?
- e. In what ways do teachers and district staff develop their understanding of assessment?

2. Understanding the Program Impact in Schools

- a. What practices are taking place with teachers and Area/district meetings with school personnel around the use of the Benchmark Assessments? To what extent do these adult participants believe that they are using the Benchmark Assessment process in ways that impact student experiences and learning? What’s happening...what are they doing?
- b. In what ways do teachers and district staff develop their understanding of assessment?
- c. What issues must be considered and addressed in planning to scale up the program under either of the two formats?

The data for this report comes from observations of 12 professional development sessions for the Math Benchmark Assessment program, 13 focus groups and 21 interviews with teachers, school-based specialists, Area Math and Science Coaches, and school principals regarding the implementation of the Math Benchmark Assessment program. The professional development sessions observed include those for teachers/Specialists from both pilot Areas and those for OMS Facilitators, City-wide Specialists, Area Coaches in addition to the planning meetings between OMS/LITD. Additional data collection took place at two schools that were piloting the program, one from each of the Areas. If we think about these data in terms of the three rounds of administrations of the assessments, the data could be classified as in Table 1 below. Each administration of the assessments included a planning meeting with LITD staff, OMS staff, Facilitators, and City-wide Specialists who helped run the professional development sessions. Each Area had a professional development day (or multiple ½ days) for the teachers or the trainers. After each, planners met to discuss these sessions; however, we did not attend any of these “post” meetings. Although we did not observe the administration of these assessments in our case schools, we did observe a Specialist in one of our case schools distributing assessments to grade level teachers. After administering and scoring tests, schools received reports from OMS/LITD. We also attended one feedback session that OMS provided to Principals and Specialists/Lead Teachers regarding the scores on the assessment. In Table 1 we mark with a “X” those events during the rounds of the assessments when we, as external evaluators, collected observation or interview data.

Table 1

	Planning	PD for	Post	Admin of	Scoring	Reports	Feedback
--	----------	--------	------	----------	---------	---------	----------

	Meeting	Area	Meeting	Assessments		to schools	Sessions
Round 1		X		Oct. 24-27			X
Round 2	X	X		Jan 18-26	X		
Round 3		X		April 26-May 4			

Findings

The findings are divided into four major sections: 1) Description of professional development sessions; 2) Exploration of how teachers and trainers make sense of and understand project goals and the purposes of assessments; 3) Understanding the two models of professional development; and 4) Understanding the program impact in schools. This manner of sharing findings draws on the evaluation questions initially framing the study and organizes the findings in a manner to most concisely offer project leaders information they can consider using in future implementations of assessment systems. Section 1 includes information pertaining to evaluation questions 1.a. Section 2 addresses evaluation questions 1.b., 1.e., 2.b. Section 3 deals with evaluation question 2.c. Section 4 covers evaluation questions 1.c., 1.d., 2.a., 2.b., 2.c.

1. Description of professional development sessions

The Center for the Study of Learning, Instruction and Teacher Development at the University of Illinois at Chicago and the Office of Mathematics and Science led three duplicate sets of professional development sessions on either restructured/in-service days or on Saturday mornings and/or afternoons for Area 9 and 12 teachers on the Benchmark Assessments. The professional development sessions were organized by grade level and structured to provide participants with an understanding of the purpose of the Benchmark Assessments, the content these assessments were measuring, the method for analyzing and scoring these assessments, and time to score student work. Professional development instructors consisted of Math Facilitators, City-wide Specialists, Area Math and Science Coaches, LITD staff, and OMS staff. These instructors also met pre and post the professional development sessions to plan for and review the sessions.

Generally, professional development sessions contained opportunity to learn about and discuss the assessment rubrics and the anchor items, opportunity to score student work, and opportunity to reflect on the use of the Benchmark Assessments. In each of the three rounds of professional development, time for learning about and discussing rubrics and anchor items was double that given to teachers to work on scoring student work. For example, in the first round day-long professional development, two hours were spent on the anchor items, a half-hour on reflection of how the assessment could be used by teachers to guide their instruction, a half-hour on scoring sample responses, a three quarter-hour on scoring their own student responses in pairs, a half-hour to discuss multiple choice reports, and a half-hour to talk through next steps. In the first and third rounds of the professional development sessions, participants were given time in the session to reflect on their practice or their use of the Benchmark Assessments. For example, in the first professional development sessions, OMS staff led participants through a reflection on the how they might use what they learned from the Benchmark Assessments while in the last sessions, participants shared their reflections of

the Benchmark Assessments pilot program and the use of Benchmark Assessments as a tool to guide their instruction with external evaluators.

The majority of time in professional development was given to discussion of the assessments themselves: the formats and rubrics of extended response, multiple choice, or short answer questions and/or scored results from these assessments. Included in this discussion was deep conversation about anchor items. In some instances, the discussion was on the mathematics lesson involved within these questions; at other times, how to appropriately score such items. During an October professional development session, for example, teachers were divided into grade level groups to practice scoring example answers to extended response questions. In one grade level group, a LITD staff member provided the group with two anchor (example) responses to exam questions and explained how and why the response was graded. The LITD staff member encouraged group members to think through the provided rubric as they looked at the examples. While reviewing anchor scoring, she told the teachers that they should consider the instructional values of scoring (i.e., the value of skills applicable to the classroom curriculum). Often, both OMS and LITD staff suggested that teachers take the conversation further and use students' responses as a tool to engage in "deep, rich conversations" with their students. For example, in a December planning with OMS, a LITD staff member explained, "the difference of 4 and 3 is not as important as the teacher understanding what is going on with respect to student understanding."

How did the sessions model standards for good practice in professional development? Overall the professional development sessions across the year allowed teachers and trainers time to have discourse around challenging intellectual ideas, apply some of these new ideas about assessment, reflect on practice, actively engage in work during the sessions, and express their views as experienced teachers regarding the usefulness of these assessments. These are all critical facets of good professional development that were evident in the sessions observed.

Teachers and trainers also expressed that they particularly valued having the chance to work with teachers from other schools who attended the professional development sessions with them. They learned from these other teachers who were in their same grade levels at different schools. Some of their comments (from interviews, written reflections, and focus groups) on this facet of the professional development sessions included the following:

- We liked evaluating together as a group and seeing what the answers were and how we got to them. Everyone benefited from that [chorus of agreement].
- Helpful because it gives an opportunity to talk with other teachers about the process.
- Some PD sessions are good. Practice grading and communicating your thoughts with other teachers/facilitators was useful.

However teachers expressed also their concerns with the professional development sessions. Some teachers regarded the "train-the-trainer" model as problematic because this model did not train all of the teachers. Said one teacher representing this view:

- Teachers training teachers is unacceptable because it only gives people who did not attend the session an overview of what was covered—no practice/examples.

Teachers especially disliked the redundancy of professional development sessions:

- Redundant. Same information over and over again. Teachers are carrying the burden. They are forced to come to PD, even though they may feel strong and knowledgeable. If teachers cannot answer the questions or have the wit to ask a colleague, then they have bigger problems than PD can address. When attendance at PD is held over teachers' heads by principals, they are not getting the benefit of the sessions. It is combative.
- Once you learn the rubric and accept the fact that grading is subjective, you do not have to be walked through everything.

2. Exploration of teachers' understanding of project goals and the purposes of assessments

In initial professional development sessions, the instructors aimed to identify and clarify the purposes, goals, and content of the Benchmark Assessment. Prior to the October assessment, staff from the Office of Math and Science reviewed the purposes of the Benchmark Assessment with specialists in a September meeting. In a PowerPoint presentation, OMS described the Benchmark Assessment as a formative assessment designed to assist and provide an intermediate measure of student learning. The goals of the Benchmark Assessment were threefold: 1) Support teachers at the classroom level, 2) Support professional development in the areas of mathematics assessment literacy and the linkage of assessment to mathematics teaching and learning, and 3) Promote discussion among teachers, students, principals and other staff about student responses. After listening to the presentation, specialists reviewed the Illinois Learning Goals and Standards that corresponded to items on the Benchmark Assessment. Specialists at this October meeting shared with OMS and LITD staff that they felt that it was important that teachers and administrators understood the purposes of the assessment, as described in the PowerPoint presentation. This sharing of the goals continued throughout the enactment of the Benchmark Assessment. The CMSI website contains a Q & A about the benchmarks and even as late as the last professional development session in May, instructors were reminding teachers of the goals and purposes of the Benchmark Assessments (See *Benchmark Assessment Fact Sheet* and *Frequently Asked Questions* at <http://cmsi.cps.k12.il.us/ViewProgramDetails.aspx?pid=157>).

a. Views on project goals.

What did teachers, specialists, and principal view as the purposes of the Benchmark Assessments project? Based on interviews, focus groups, and written reflections, we found that they perceived that the Benchmark Assessment intended to address four goals: i) Assist teachers in obtaining a detailed understanding of students' mathematical knowledge, ii) Facilitate changes/adjustments in instruction, iii) Prepare students for the ISAT, and iii) Provide teachers with opportunities to talk with other teachers. The following data (words of teachers and trainers from focus groups and written reflections) are representative of their views on these goals:

i. Assists in understanding students' knowledge of math

The Benchmark Assessment could inform teachers with a more detailed understanding of student knowledge of math, as well as student performance in math:

- The Benchmark Assessment is an assessment tool for teachers to gather information on the way students are performing.

- It's informative to know what our kids know and what we need to spend more time teaching.

ii. Guides instruction

Specialists and teachers similarly regarded the Benchmark Assessment as a tool to guide instruction:

- The Benchmark Assessment is a formative assessment that is supposed to get teachers to reflect on instruction.
- Encourages differentiated instruction.
- Low or no stakes test to guide instruction.
- Pilot assessment to improve teaching strategies/use the data to improve math curricula and skills.
- To help teachers identify strengths/weaknesses of students so that teachers can zero in on those areas & change strategies to address them.
- Opportunity for teachers to review and refocus their practices—What do teachers need to work on?
- Encouragement for teachers to move away from more traditional teaching strategies.

iii. Preparation for the ISAT

Some school staff perceived that the assessment served to prepare students for the ISAT:

- [The Benchmark Assessment] is used to guide teachers toward better instruction especially for Extended Responses and the ISAT.
- To prepare kids for the ISAT test—the format is similar to what they will see on the ISAT.

iv. Opportunity to talk with other teachers

The assessment could also provide teachers the opportunity to talk with other teachers about particulars of the assessment as well as student learning:

- Opportunity to talk with other teachers about the process and results, which is particularly useful when thinking through scoring).
- OMS has said that the purpose of the assessment is to get teachers talking about learning.

b. Understanding scoring, students, and assessment

This section details the issues/challenges that arose throughout the year during professional development and how professional development instructors addressed these issues/challenges. First, both teachers and OMS instructors experienced difficulties with achieving consensus in their scoring of practice questions. LITD/OMS instructors responded to these difficulties by encouraging teachers/Specialists to focus on understanding student thinking rather than concerning themselves with the test scores. As part of their training model, LITD/OMS instructors also deliberately engaged teachers in scoring examples of generic student assessment responses rather than their own students' assessments. When teachers/Specialists complained about this focus on generic student work, LITD/OMS instructors maintained the importance of learning about the process of scoring. Next, teachers/Specialists and OMS/LITD instructors

frequently compared the assessments to the ISAT. Although OMS/LITD instructors emphasized some of the differences between the ISAT and the assessments, teachers/Specialists found it difficult to differentiate the different types of assessments. Throughout the year, OMS/LITD instructors reinforced the message that the Benchmark Assessments, unlike the ISAT, were a tool to guide instruction.

i. Scoring - Difficulties of achieving scoring consensus

At the October professional development session, a LITD member anticipated problems with the upcoming November professional development because teachers expressed concern about the gray areas in scoring (a score of “1” could arguably be a “2” to different people) and the test in general. Scoring issues did in fact emerge during the November professional development sessions when teachers practiced scoring the extended response questions. When asked to share their scores on practice questions, teachers shared different opinions regarding the importance of achieving consensus on the scoring:

OMS staff person: I know you are engaged in some rich discussion, and it is hard to stop, but let’s share what your group concluded about the score for the math questions. Let’s start with this group.

Group A: Four though we have not come to consensus.

OMS staff person: How about this group... [Various groups respond.] 3, 3, 3, 4
How about strategic knowledge? [Various groups respond.] 2, 2, 2, 2

OMS staff person: As you can see we do not come to a universal consensus. Is this a problem?

Teachers: Yes.

Teachers: No.

Similarly, at a December planning meeting, when LITD/OMS conducted a training session for OMS/LITD instructors regarding scoring of the Benchmark Assessment, participants in this session admitted difficulties with achieving consensus in scoring. The following example depicts the difficulties of scoring the example questions, as well as the need for conversations around the questions:

Facilitator A: Looking at a bar graph is a strategy...but ...is the correct strategy/answer...I can see the mathematical knowledge. I see the word least and most, but then...the second one refers to the graph. What are we using as the base of a strategy? If they don’t have it included in their answer...then they don’t get the score?

Facilitator B: They have to refer to it (the graph).

Facilitator A: So if they answer ___ that’s mathematical knowledge because they are referring to the graph...

LITD staff: Specifically note the graph and then 2nd reference to most and least...Can't not have a graph here and not reference most/least?

OMS staff person: We don't know why they chose what they chose...saying least likely doesn't say how he got that.

Facilitator A: I understand what you are saying...but the justification is going to be hard...

ii. Importance of understanding student thought processes

Following the discussion about consensus in scoring at the November professional development, OMS/LITD instructors discussed the importance of understanding student thought processes through conversations about the questions as opposed to achieving consensus on the scores. During a discussion of scoring, one of the instructors explained how student responses revealed information concerning students' understanding of mathematics.

LITD Instructor: What is the purpose of the activity? It is to develop a common understanding of what our students know, and how we judge that knowledge. The bubbles on the sheet give you an example of the thinking behind how they have come to their assessment number. This isn't the absolutely right score, but your answers were around it.

OMS Instructor: When the ISAT committee gets together, it is about a group of ten teachers. They have to have consensus on their scoring. We do not have to have that level of reliability. We want teachers to use this to impact instruction. Teachers need to know that there is a range of scores.

At multiple professional development sessions, teachers participated in the "OMS thought process" as they scored examples of generic student test responses rather than the actual tests done by their students in order to facilitate group discussions around the scoring rubric. At the January 23rd planning meeting between LITD and OMS, an OMS staff person emphasized scoring in the agendas for the February professional development sessions. That is, teachers would have some time in the early morning to reflect on the January administration of the Benchmark Assessment. The later part of the morning would include discussion and scoring of the short answer questions, as well as of the anchor extended response examples practiced by LITD/OMS instructors. OMS staff also encouraged the Facilitators and City-Wide Specialists who were instructing the February professional development sessions to explain the scoring of the extended response items. OMS staff hoped that the instructors would look at the rubric and share with teachers the "point of the problem" (i.e., what is expected for good mathematical and strategic knowledge).

In fact, at a February professional development session, after dividing the teachers into groups to analyze a test example, the instructor advised the teachers to "take the knowledge gained from the practice (or scoring examples) and use it when scoring their own students' work." Teachers hoped to score their own students work at professional development and expressed frustration with this focus on student examples. For example, one teacher complained, "There's no point of going through the terminology of this rubric because it is not specific to our students' work. An OMS Facilitator responded, "Before you can get there, you have to

figure out how to judge. We will look at student work soon.” As the professional development sessions continued throughout the year, teachers continued to spend more time on learning about and discussing rubrics and anchor items and the mathematical and strategic knowledge inherent in them than actually scoring student work.

ii. Tensions in distinguishing the Benchmark Assessment from the ISAT

In addition to scoring, another issue/challenge that emerged at professional development sessions for the Benchmark Assessment concerned the comparisons between the Benchmark Assessment and the ISAT. Tensions existed in conversations about the Benchmark Assessment vs. the ISAT as OMS and LITD desired to distance the Benchmark Assessment from the ISAT but also recognized the similarities between the assessments. Comparisons between the Benchmark Assessment and the ISAT emerged in the initial professional development sessions and continued throughout the year at other professional development sessions. Some teachers expressed confusion about the differences between the two assessments and referred to the Benchmark Assessment as a *pilot ISAT* in an October professional development session. One OMS staff member, who was worried that teachers would not understand the purpose of the Benchmark Assessment because of its similarities to the ISAT explained to LITD staff, “I think that the message to teachers has to be clear... it looks similar to ISAT and I know you aren’t telling them it is for ISAT...but so much is embedded...we are expecting a higher level...”

It is possible that the comparisons between the Benchmark Assessment and the ISAT occurred naturally since the rubric used to score the extended response questions on the ISAT was also used to score the Benchmark Assessment. An OMS instructor at a November professional development session explained to the teachers, “We use the ISAT rubric to get an understanding. We want you to see the complexity of scoring and what the students need to understand.” At the November reporting out session with OMS and principals, teachers and Specialist, the Office of Math and Science addressed the relationship between the ISAT and the Benchmark Assessment by saying they were “using (the) ISAT scoring but not (the) ISAT process...Benchmark Assessment is all about using the scoring for instruction, not grading.”

In conversations at professional development, as well as in focus groups, teachers raised a variety of different concerns about the similarities between the Benchmark Assessment and the ISAT. One teacher, for example, felt that her students were at a disadvantage with the Benchmark Assessment because they had never taken the ISAT:

Teachers who hadn’t given IGAP or ISAT before were at disadvantage because they hadn’t made that adjustment. Teachers who have had experience of doing ISAT for a few years had already changed their long-range plans and had moved all of those topics up to the front of the year. It was very anxiety provoking for the children. Because some kids...with an Iowa type of basic skills, they plummeted. And the kids and the parents were really concerned. They couldn’t understand, and, actually, it was really teaching almost a different subject the way it came about—because it’s all higher order thinking skills versus concrete knowledge. 7th graders never had the ISAT so it was a whole different kind of ...The 8th graders did have the ISAT.

Another teacher at a November professional development session was concerned that the Benchmark Assessment resembled the ISAT too closely and equated the Benchmark Assessment with the ISAT: “I think there is a real issue to use the same kind of test that the

state issues.” The professional development instructor addressed this teacher’s concerns, “It is not a practice ISAT. The [Benchmark Assessment] is for instructional purposes. It is to understand what they know about that stuff. It is not supposed to be an exact replication of the ISAT.” In a May focus group, a teacher noted that “the fact that [the Benchmark Assessment] imitates the ISAT and we score [Benchmark Assessment] like the ISAT has the unfortunate effect of making the Benchmark Assessment guilty by association with testing that has strong emotions attached to it.” Not all of the comparisons to the ISAT were negative. One teacher, for example, felt that the Benchmark Assessment could prepare students for the ISAT; “The BA is a helpful reminder to teachers to use the language that may be on the ISAT so that students are acclimated to it.”

Even in February professional development sessions, OMS staff continued to emphasize the differences between the ISAT and the Benchmark Assessment, which ultimately connected to their goals for the Benchmark Assessment as a tool to aid instruction. For example, an OMS instructor described the Benchmark Assessment as “a formative test—not like the ISAT.” She also reminded the teachers that the Benchmark Assessment should only be used to inform practice. She explained that OMS added short response questions to the BA to align the test more closely with the structure of the ISAT. For some teachers, the differences between the Benchmark Assessment and the ISAT created confusion. For example, two Area Coaches explained how teachers were confused by differences in the format and scoring of the questions in the two tests:

Coach: The extended responses were not done in the format that they are presented on the ISAT, so that threw the teachers off when they gave the extended response. Then what happened at the PD in the conversation was that when it came to scoring them, you wanted to give them a knowledge, a strategic and an explanation score for one scoring number for each problem. Well, if you had a problem that had six different parts, the teachers came and said, you know, ‘each part can be scored individually; each part can have a strategic knowledge and a plan and an explanation.’ So, then, if you do that, how do you come to one score for a problem? And, then if we are going to expose our students to this type of extended response in October, January, and March, our students are going to get used to this, and they are going to practice that way. And then when they get the ISAT in March, it doesn’t follow with that. So where they are learning how to break it up and do parts of it and explain each part, that is not what is going to be tested in the ISAT, so it’s going to be confusing.

Although OMS planned to revise the questions, the questions would still consist of different parts:

Coach: But we were told at the principal’s meeting that even though they were changed, that there would still be some with parts—the students have to be exposed to all different types of extended responses. So there will be maybe one with no parts, and there may be some with parts.

Despite these comparisons to the ISAT, at all of the professional development sessions, consistent with the initial goals presented, instructors emphasized the potential for the Benchmark Assessment to inform instruction as opposed to high stakes testing. Professional

development instructors spoke about the Benchmark Assessment as a tool to aid teachers, to guide their conversations with students, as well as their classroom instruction.

3. Examining two models of professional development

Implementing the two training models represented another issue/challenge of professional development. As mentioned in the introduction, in one Area, specialists or lead teachers were trained in the Benchmark Assessment (i.e., “train the trainer”); in the other Area, teachers received training (i.e., “train the teacher”). At the September 23rd staff meeting between OMS staff and Specialists, an OMS staff person described the purposes and goals of the “train the trainer” model; “The goal is to have teachers score the responses and share this information with their teachers. One teacher from each school would go to the training and be the point person for that school.” At this same meeting, OMS planned to incorporate the naturally occurring interactions between Specialists and teachers when designing professional development sessions for the Benchmark Assessment; “We need to have the support, not the passive support, but the active support from the AIOs. We need to understand how you [Specialists/lead teachers] conceptualize sitting down with teachers, and if there is a pattern already, then we need to know that so that we can infuse that in our training (OMS staff person).”

As the year progressed, OMS actively sought information to better understand how these training models worked, especially as they planned for the use of the Benchmark Assessment in the following year. For example, at a November feedback session, an OMS staff person desired to learn from principals, lead teachers, and Specialists “What’s a viable model for 500 schools?...What happened in your school when people came back to your school is critical to us...looking at huge numbers together to grade the tests...this is of interest to us.” Likewise, in an April meeting with school Specialists, principals, and Area Coaches, OMS staff continued to express their hopes for as well as the challenges of building coherence among teachers. An OMS staff admitted that OMS needed to “think really carefully about how messages at [professional development] training can get back to the school.”

Since this year marked the piloting of the Benchmark Assessment, it is worthwhile to understand how the process worked for all involved—the teachers, Specialists, principals, professional development instructors, and OMS central office staff. For example, how did the various training models work? Did teachers and Specialists really train other teachers in understanding the purposes of the assessment, as well as scoring the assessment?

By implementing the “train the trainer” and “train the teacher” models, OMS and LITD hoped that teachers and specialists who were trained in the Benchmark Assessment would share the knowledge they had learned at professional development sessions with others at their schools. As expected, the effect of the training models varied by school. At some schools, the goals of collaboration and sharing of information were not achieved. At the April 18th meeting to “train the teachers,” an OMS staff person confessed, “some teachers were having no collaboration about [the Benchmark Assessment].”

A conversation with two Area Coaches was especially revealing about the differential effects of the training models. The Coaches shared that in the “train the trainer” model, schools relied upon their Coach for assistance while in the “train the teacher” model, the Coach had a hard

time getting into schools to assist because these schools perceived that they understood the Benchmark Assessment and, thus, did not “need” assistance. Therefore, using the “train the trainer” model seemed to give Coaches a greater entry into the schools. Once in, the Coach had the opportunity to monitor and assist school staff at grade level meetings or at all staff meetings to keep teachers focused on the true tasks of the Benchmark Assessment. For example, after gaining entry, a Coach noted her ability to direct the conversation away from the individual towards classroom trends—keeping the data from being misused or misunderstood. Yet, our data from two case schools contains no mention of Area Coaches providing assistance under either model.

What factors contributed to problems with the training models? At a reporting session between OMS and principals or lead teachers/Specialists, we heard how the trained teachers’ lack of understanding about the Benchmark Assessments made it difficult for them to train others and how the process of getting test materials distributed, taken, scored and returned on time was sometimes unmanageable:

[We] had problems with training sessions...those who came had problems going back to explain to teachers...weren’t clear on process.

It seems that some felt they had not gained enough information either on Benchmark Assessment or the process they were to follow in terms of administering and scoring the assessments prior to the first administration of it. Some explained during this reporting session that even after the six-hour training sessions, they did not feel confident to go back to their schools to train the entire staff or had difficulty juggling the administration, scoring, and returning of forms in a timely manner. Some proposed offering summer professional development sessions for the Benchmark Assessment similar to CMSI curricula professional development, especially in light of scaling up system-wide.

5. Understanding the program impact in the schools

How did the Benchmark Assessment impact teacher instruction, given that is a primary goal of the assessment? In addition, since scoring was such a big focus of professional development, what was the scoring process really like for the teachers?

a. Impact on teacher instruction

OMS and LITD repeatedly emphasized that they wanted the Benchmark Assessment to influence classroom instruction. The Benchmark Assessment differed significantly from the ISAT because of the potential for it to influence classroom instruction. In written reflections, focus groups, and interviews, teachers described the difficulty they had with incorporating the Benchmark Assessment into their everyday instruction. The following comments represent the difficulties experienced by teachers:

The most challenging part of the process is figuring out how to use the data. Interpretation and application of the data is problematic.

That kind of information can be very useful. Speaking for myself, trying to balance the pacing of what is in the test and providing reflective time or teaching time can be extremely frustrating to me as a classroom teacher.

[The Benchmark Assessment could be] potentially more useful than it's actually been. Trouble is that one of the things that happens is that the test covers a number of topics—some have already been addressed; some are topics a teacher may be planning to address; others are topics a teacher probably never will address. Theoretically, for this to be most useful, the teacher will have to look at the results repeatedly over the course of the school year. Even with this kind of information, it's probably really hard to step back from the day to day grind of preparing your lesson plan to sort of look at the results, reflect on them, and think hard about how it's going to influence your lesson plans a month from now, or a month from now realize, "Oh, I'm planning something that was on the assessment a month ago. I should go back and look at the results from the assessment a month ago." [This is] hard for a classroom instructor to do.

Why does the test provide four different extended responses? These options make it difficult for teachers to figure out which students mastered or need help on what skills. The format necessitates additional work after the test has been administered. Teachers have to provide photocopies to all students and figure out who answered which of the four questions. Then, teachers have to use the data to work on the four areas.

The views of one Coach echo the teachers' sentiments. For example, this Coach explained to us that use of the CMSI curriculum made it difficult for the teachers to modify their instruction based on the assessment:

Coach: And I think the statement that you made previously said how you think the benchmark or the assessment should move instruction, and, see, I have a hard time seeing that because the benchmark, if you are using the CMSI approved curricula, it cannot move instruction because you are almost locked in to following their procedures—the lessons.

These sentiments were repeated at the April 18th meeting with OMS and school coordinators of the Benchmark Assessments. For example, one Specialist noted that her teachers were frustrated that they could not re-teach concepts students missed in the Benchmark Assessment because of the spiraling nature of CMSI curriculum. This Specialist shared that her teachers "wanted to immediately re-teach [concepts]" and "felt like they let [their] kids down" since they taught the concepts yet students did not master them. According to this Specialist, the teachers questioned, "Why would we get these [Benchmark Assessment] results back if we weren't going to re-teach?" Ultimately, the teachers at this school "felt like they let kids down" and since there was no growth school-wide, this also "kind of made them feel like failures."

Although LITD/OMS instructors emphasized the connections between the Benchmark Assessment and classroom instruction, they did not provide teachers with many strategies for incorporating what they learned from the assessment into their instruction. For example, one Coach explained that the instructors "are not telling us any strategies; at least I am not getting any strategies from them. The strategies that are going to be used in the classroom are all coming from the discussions of the teachers." Similarly, another Coach shared that OMS discussed Benchmark Assessment "in terms of the CMSI curricula; the message has always been stay with the curriculum; keep with it exactly as it is without any supplemental or without any other suggestions of strategies to do when they are not there." Rather teachers

developed their own strategies for incorporating what they learned from the Benchmark Assessment into their instruction. For example, after a Coach repeatedly told her teachers not to “interrupt the math curricula as it is supposed to be done,” the teachers and the Coach discussed how to assist struggling students by going “back to the curriculum that they are using and putting up some centers that they can go into during free time, maybe pulling those students aside at another time but never to interrupt the math program or the math lesson of that day.”

b. Impact on teacher scoring

OMS staff admitted to problems with scoring in their April 18th meeting with school Specialists; “One of the big lessons [learned was it is] really difficult to find time to do [the] scoring. Teachers end up doing a lot of scoring on an individual basis. It should be an opportunity to have discussion and collaboration.” Although OMS hoped that Benchmark Assessment “would promote discussion” among the teachers, many problems occurred in scoring. In fact, principals and specialists at this meeting expressed their concerns with the scoring process. They were especially concerned about the time involved in scoring the extended response questions, which represented only 10% of student’s scores on the assessment. The opinions of people attending this meeting reflect their concerns with the subjective nature of scoring, the time it took to score, and the roles of the Specialist and the teacher in scoring:

I do not agree with the current method of scoring the Extended Response questions. Teachers have different levels of understanding, which causes scoring to vary greatly.

Scoring took up a lot of time and energy in this year of many assessments. Teachers had to use their prep periods to score.

Scoring was a mess at our school. The first time the math Specialist tried to score as many as possible. It was overwhelming and extremely time consuming. The second assessment the teachers scored their own, which affected the data because they might have scored differently than the Specialist. And many teachers didn’t return all their assessments.

In written reflections, interviews, and focus groups, teachers, principals, and Specialists described challenges of scoring such as the time it took teachers to score the assessments. For example, one teacher commented on the lack of time they had for scoring:

Our problem is time. If you had time on site for teachers to sit down and go through and do the curriculum and make the small list, and, you know, it would be ideal, but we don’t have time. And in scoring, that’s time. And we don’t have enough on site time to do the job we would like to do.

As they scored the Benchmark Assessments, teachers also experienced difficulties in achieving consistency between the test results versus their understanding of student knowledge. The following comments represent teachers’ concerns with the accuracy of information provided by the Benchmark Assessments:

But, I know, for the multiple choice, it helps drive instruction. If I see some of the students didn't do well with a particular question, you go back and re-teach for mastery. So I found those to be helpful. But extended response, I found that to be, like the information they send back, I have contradictory type issues with it. Like they come out with 4-4-4, I don't agree with their scores. So I question, "Am I doing the right thing? Is that the way I should grade my students too? Or is that not the right thing that I have to change? Because I have seen some that have been perfectly scored that I would not have given them the same score, so..."

If your school is on probation, we don't have that luxury of saying, "Well, fine, this kid's got a 2. This kid's got a 3. Fine." We don't have that luxury because if we are going to/for the highest stakes because your school, and your [life?] depend on it, then you need to know. Am I beating a dead horse because this kid really got a 4? And I'm saying that this kid got a 4 so I'm spending time here trying to guess what I consider is a 4. Or am going to the other extreme? I feel this is a 4, and it's really a 2 or God forbid it's a 1! But I'm seeing, "Oh the kid said this, the kid used these terms, the kid did this, and, therefore, I'm going to give the kid this 4, and it's erroneous. So in that sense it is very important that we understand how it is being looked at, at the state level. We need a consistency and a consensus because we may come to a consensus that this is a 3-3-3, it goes through the state where it has the 1-1-1. Well, you're up the creek without a paddle, and [going in] the wrong direction. So this is part of it. I think one of my frustrations is trying to [balance the two]. I know how to assess my children, okay? You give me a half way decent tool...I have the tool, I cannot use it as effectively as I would like given the constraints that I'm under.

Doesn't provide accurate information about the students. Straight A students who are always getting...They take this and then they get 1's. And that doesn't tell me they don't know anything. It tells me either they didn't get this question, it was too confusing, or by the 3rd time they didn't care. And I know that came up. After they're done with the big test, and I say, "You have to take this again." I mean, they're done in 2 minutes. They don't care. So I don't think it's accurately providing much information about the students. Like I wouldn't base everything on the first test, or we'd need to go back to learn how to count. Maybe they didn't feel like counting that day because they're tired of taking tests. Maybe they didn't know they needed to count because the question is confusing. Or they didn't read the directions because as you said it's the same question as the last one.

Despite these challenges, teachers appreciated talking with other teachers about the scoring process (e.g., "opportunity to talk with other teachers about the process and results [which] is particularly useful when thinking through scoring.") and felt that they learned more about their students through scoring (e.g., "I think the experience of scoring the extended response really, really makes teachers hone in on the individual student. Makes us more aware of what our children are doing. Just the time scoring the individual response. It gives you insight into their thinking. You know if they haven't a clue. If they have an idea, but just can't express it."). The following comments represent teachers and other school staffs' appreciation of the information provided by the Benchmark Assessment:

Teachers can grade them in their grade level meetings and see how students are performing. They'll know what to teach. Teachers don't mind administering the tests, and they do coincide with Reading. School does in-house assessments.

Better able to group students for specific issues, such as geometry problems, based on discussions analyzing these tests.

Third assessment is more useful than the first two because some of the material had not been covered [in previous administrations]. The last assessment helped teachers see the big picture, whereas the first two threw the teachers off.

Helpful to see how they were doing on the extended response—were they writing at all, what they were writing.

How did the Area Coaches and school-based Specialists participate in the scoring process? The roles of Area Coaches and school Specialists in the assessment varied according to each school. At some schools, the school-based Specialists were responsible for scoring. For example, one Specialist explained, "I did all the scoring, which wasn't bad but didn't give teachers the opportunity to look at the work prior to being scored." In comparison, one Area Coach found it difficult to assist teachers with scoring of the assessment as entry to the school for this purpose seemed to be blocked (at least for the first administration of the assessments) because all teachers had been trained. In addition, although teachers were doing the scoring at one school, the Coach explained, "There is no follow up [with the math Specialist]." However when Coaches were given entry into schools, they found that their presence helped carry the use of Benchmark Assessments to a deeper level. Coaches offered a few examples:

Coach: One of my teachers (and this happened at an implementing school) they have MTB, she said that she couldn't deviate from the curriculum, and she said but I could mark where it's coming up and that's where they [teachers] started the conversation. And so then, it came from a teacher and so she asked "Could I do that?" and so other teachers started pulling out their textbooks. And they started looking and saying, "Ok, it's here" and putting a post it note, and someone said "Let's write down what the distractor means." So this is what they came up with because the teachers are talking about it so the value in this is this assessment was in the conversations that were engendered in the teachers while they were talking about student learning. But this direction doesn't come from elsewhere...These conversations, though, would not have taken place unless [we] were at those schools....Because [we] were there, we are able to make sure that the emphasis is not on the individual—It's on classroom trends—so we are able to direct the conversation to make sure that it goes in a place, it doesn't go to a place where the data is going to be misused or misunderstood.

Coach: And the other thing that I am really trying to stress to the principal is because they have now gotten into with the Learning First where they are getting this information and they are grouping their students by skills that are missed for differentiated instruction so they want to use this benchmark math assessment in the same way where they are going to say, "Well, all of these students missed this skill so let's put them together and teach this skill." And what we are trying to say is "No, we don't want you to do this. This is not the purpose of the test. We don't want you to

deviate from the curriculum. We want you to continue to teach math, but if it has been taught and the curriculum has marked it as a mastered item then we have to make some sort of [effort for students who have not mastered it to pick it up]”

Recommendations

This section of the paper describes recommendations to consider for improving next year’s Benchmark Assessment professional development and administration in schools. We draw on interviews, focus groups, and written reflections with teachers, specialists, and principals throughout the school year, especially at the April 18th meeting and the May 6th and May 13th professional development sessions on the Benchmark Assessment. While we have learned from specific recommendations that those we spoke with made, these recommendations are our own—drawing on multiple points of data and our analysis of it in context.

1. *Around professional development*

Get all of the right people to participate in professional development workshops

Based on the above findings, it was clear that CPS needed to find better ways to get the necessary teachers and administrators to attend the professional development sessions. When teachers raised this issue in written reflections they noted that they found the “train-the-trainer” model “unacceptable” and wanted all teachers from their school to attend so they would get a better understanding of the project so they could have “deeper conversations” back at the school. Further, teachers wanted to see the principals attend sessions and enforce some accountability so that the project was used effectively at their schools. Having more staff from each school would also foster “buy in” for the project.

What motivations would get more of the right people to these workshops? Some teachers suggested they be mandatory. Key to getting the appropriate teachers to participate with enthusiasm is how math Specialists, OMS, Areas, and the principal promote this program in the schools. Getting these leaders to buy into and promote the program is the critical first step.

The timing and location of the professional development sessions could also be improved to try and foster ease of attendance. Some of the teachers offered various alternatives like

- In-school professional development sessions, rather than Saturday training.
- Scheduling sessions at convenient locations with ample parking.
- Giving more notice about upcoming sessions.

In addition, perhaps Area Coaches, City-wide Specialists, and OMS Facilitators can strategize ways of getting schools they work with to have joint professional development and scoring sessions as a way to both alleviate travel time and increase the opportunity of building professional community around math instruction across schools.

Utilize the enjoyment teachers have in interacting with each other

Teachers identified their interaction with each other around the assessments as both a goal of the project and one of the most enjoyable parts of their participation. CPS would be well

served to play on this strength of the project. This could be another means to fostering additional attendance at workshops. If teachers knew that the majority of their time at workshops would be engaged in productive conversation with peer teachers, this could boost attendance. One teacher suggested that teachers do their scoring of assessments ahead of time and spend more of the workshop in dialogue and “trading ideas” about the student work.

2. Around administration of the Benchmark Assessment

Provide better information so teachers know how to use benchmark assessments

The teachers expressed their need to better understand what strategies to use in order to apply what they learned from the assessments to their teaching decisions. They wanted to understand what to do once they had the scores for their students. They wanted to understand the logistics around how to apply the assessments to their classrooms.

One clear challenge many teachers faced, as we note in the findings above, was the issue of how to reconcile the opportunity to use the assessment scores to better serve their students with the message they had received about CMSI math curricula they were using which spiraled and should be taught according to plan without deviation. We believe that OMS leaders need to talk through this issue—those supporting the Benchmark Assessments and those supporting the CMSI curricula—and come up with guidelines and illustrations for teachers on what good solutions might look like. They would then need to clearly communicate these options to all schools and the Specialists, Facilitators, and Coaches who support those schools and these CMSI programs.

Teachers and other school staff proposed a variety of other suggestions to improve their ability to use the Benchmark Assessment for next year. These suggestions fell under the categories of themes of timing, the content of the assessments, and support resources for the project. Suggestions on ways to changes the timing of the test and time provided to score the tests included the following:

- Change the timing of the test: October is too early.
- The math assessments and Reading First test should be given at the same time like the ISAT.
- Make sure there is built in time for scoring...[and] to discuss the results.
- Please provide time for teachers to score the extended response on site and even more importantly, provide time for data analysis and planning strategies to utilize the data.

Suggestions on possible changes in content/questions included

- The length of the multiple-choice assessment was not sufficient. It should have at least double the number of items. Maybe 25 questions instead of 12.
- Test them on what they’ve been taught.
- Test should include more extended response questions.
- Teachers and students that are using CMSI curricula are probably better prepared to handle the BA. The wording of many lessons in Math Trailblazers, for example, force teachers to use the language that will be on the assessments, and, in turn, triggers

students to think about them in a particular way. OMS should track who is using which curricula and study how the students perform... Teachers who are not utilizing CMSI curricula should not be at a disadvantage when working with the Benchmark Assessments. Regardless of the curricula, the test should cover the same material.

- The vocabulary used in the test is problematic. Often, the students do not understand words that, if explained, would not give the answer away, but would help them understand what the question is asking. For example, if the question is going to include the word “baseboard,” then perhaps it should also include a picture of a baseboard. This would not make the test question easier, but at least give them a chance to figure it out.
- Test questions should include an explanation of any potentially difficult vocabulary that is neutral to solving the problem.

Suggestions on changes in resources provided included

- Teachers should be provided with a booklet of sample extended responses.
- Should provide a different rubric for each problem.
- Provide instructional booklet so teachers can meet after school and score together. If have rubrics, (not reliable) since different teachers come up with different scores for the same test. Don’t think it’s valid. Investing lots of effort on scoring something that is only worth 15% of the total score—ISAT only 5% Extended Response, 10% Constructed Response.
- Principals should have a mathematics coordinator who can focus only on the Benchmark Assessments and monitoring that the teachers are implementing the skills and concepts needed for extended response and multiple choice questions.
- Time should be allotted, in school staff development to just analyzing the data among all teachers.

Deal with issues that confuse teachers with some concrete efforts

The messages that teachers received about how the Benchmark Assessments related to the ISAT left them with some confusion. CPS needs to address teacher ideas about the relationship or lack thereof between these different tests. A document that outlines talking points for OMS, LITD, Area, and school leaders could directly address the similarities and differences between Benchmark Assessment and ISAT. This document could pose some of the conceptions and misconceptions about tests. It could also offer some examples of correct ways to understand the Benchmark Assessments related to ISAT and incorrect ways.