Lessons Learned about Videotaping and Coding Classroom Observations

Teresa Garcia Duncan

Rosemarie Ataya O'Conner

ICF International

Presented at a symposium, "Math for All: Lessons Learned from Piloting a Randomized Controlled Trial in a Large Urban District" at the annual meeting of the American Educational Research Association, Washington, D.C., April 2016. This research was generously supported by a grant from the U.S. Dept. of Education, Institute of Education Sciences (Grant No. R305A140488 awarded to Education Development Center). Opinions expressed in this paper are the authors' and not of the U.S. Dept. of Education. Please address all correspondence to: Teresa.Duncan@icfi.com.

The authors gratefully acknowledge the contributions of Aikaterini Passa, who helped with data analysis, and of Kelle Falls, who helped manage and implement the data collection and coding process. This work would not have been possible without the efforts of Alyssa Carr, Jean Dauphinee, Elly Field, Lisa Luo, Erica McCoy, Sarah Pfund, Kasia Razynska, and Jessica Zumdahl, who collected data and coded videos.

Introduction

Professional development (PD) that is focused on teachers' knowledge of academic subject matter and how students learn that content has been found more likely to be related to changes in classroom practices and enhanced student outcomes than traditional approaches that focus mainly on the processes for delivery of instruction (Cohen & Hill, 1998; Corcoran, 1995; Garet, Porter, Desimone, Birman, & Yoon, 2001; Kennedy, 1998). A small number of PD programs – in particular, Math for All (MFA) – do integrate learning about how to differentiate instruction with learning about mathematics content (e.g., Brodesky, Gross, McTigue, & Palmer, 2007; Moeller et al., 2012). However, there is a paucity of rigorous studies that link PD to student outcomes (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007).

In fall 2014, the Institute of Education Sciences (IES) funded an efficacy trial of MFA to help build the knowledge base on the impact of PD interventions. A small pilot of 20 teachers and 339 students in four Chicago Public Schools (CPS) was conducted January-June 2015. The full study is taking place in 2015-16 (implementation year) and 2016-17 (impact year), involving 32 CPS schools, 256 4th and 5th grade general and special education teachers, and approximately 6400 students. Third grade test scores will be used as baseline data for 4th grade students and 4th grade test scores as baseline data for 5th grade students.

In this RCT, our research team is examining the impact of MFA on both teacher outcomes (i.e., knowledge, skill, and classroom practice) and student outcomes (i.e., academic achievement in mathematics and efficacy). Classroom observations are a key component of the MFA study. Observation data provide us with key insights into classroom instruction, the quality of teacher-student interactions, classroom climate, and allow us to directly compare the pedagogy of control and treatment teachers.

The results of the implementation and impact years will add to the knowledge base, but this paper is intended to add to another aspect of the knowledge base: the considerations and challenges involved in conducting a "real-life" RCT – in this case, lessons learned about videotaping and coding classroom observations. We have remarked multiple times to ourselves during the past several months about the wisdom of having built in a pilot into our research plan, and believe other researchers will benefit from our sharing the lessons learned (thus far). Sharing what we are learning from this RCT can help improve the conduct of future RCTs, particularly those set in large, urban districts like Chicago.

Method

Math for All PD

MFA consists of five one-day workshops and classroom-based assignments, providing a total of 50 hours of PD over one school year. The program uses video cases and a lesson-study approach to engage general and special education teachers in collaborative lesson planning to make standards-based mathematics lessons accessible to various kinds of learners. While the intervention was designed with a focus on improving math education for students with disabilities, all students are thought to benefit from instruction individualized to their specific learning needs.

Data Collection

During Year 1 of this project (2014-2015), we conducted a pilot study with a group of 20 teachers from CPS to test research instruments and procedures for data collection, recruitment, and intervention in preparation for the main RCT. We completed the implementation of the professional development for the pilot study in April 2015. Data collection for the pilot study, including the collection of student and teacher surveys, video recordings of classroom practices, teacher logs, and interviews with teachers and principals were

completed in June 2015. Based on the findings from the pilot study, we refined our instruments, including the student and teacher surveys, and the teacher logs. We also fine-tuned the coding scheme for the MFA Teacher Performance Assessment and made the decision to use the Classroom Assessment Scoring System (CLASS) instead of the Mathematical Quality of Instruction (MQI) as the primary classroom observation rubric (see discussion below). Researchers from ICF participated in training for the CLASS instrument in July 2015, and the coding of classroom videos using the CLASS and MQI (coding of the latter was conducted by consultants trained in the use of this instrument) was completed in October.

Measures

Mathematical Quality of Instruction (MQI). The MQI protocol for classroom observations, developed at the University of Michigan and the National Center for Teacher Effectiveness (NCTE) at Harvard University (Hill, 2010), is designed to measure the mathematical work that occurs in classrooms. The instrument provides separate scores for various elements of effective mathematics teaching in three different areas: teacher-content relationship (richness of the mathematics, meaning-making, mathematical practices, errors and imprecision); teacher-student relationship (working with students and mathematics, responding to students' mathematical ideas, correction of student errors); and student-content relationship (participating in meaning-making and reasoning, connections between classroom work and mathematics). The MQI scoring protocol was designed for assessing videotaped mathematics lessons. Each lesson is divided into five- to seven-and-a-half-minute segments for scoring. Raters assign each segment a score for each of the MQI elements, and also assign an overall score to the whole lesson. Each lesson is scored by two raters working independently, and scores are averaged across lessons to derive a teacher score.

Classroom Assessment Scoring System (CLASS). The CLASS measures the quality of teacher-student interactions within four domains: emotional support, classroom organization, instructional support, and student engagement (Pianta, Hamre, & Mintz, 2012). Each of the domains is divided into dimensions of classroom quality. Observers typically watch a lesson for 15 minutes, taking notes on the specific behaviors they observe related to each of the CLASS dimensions. Scoring is completed at the dimension level using a 7-point scale, with the low range being a score of 1-2, the middle range 3-5, and the high range 6-7. The CLASS manual provides detailed information to help observers determine the specific score. The observer then watches the next 15 minutes and scores each of the dimensions again, repeating this cycle of observation and scoring until the end of the lesson. Lesson scores are created by averaging scores across all 15-minute cycles, and scores for teachers are averaged across lessons. Observations can be scored live or using video.

Please refer to Table 1 for a summary of the psychometric properties of the MQI and CLASS

Procedures

Videotaping classrooms. Twenty 4th and 5th grade teachers at four CPS schools participated in the pilot.

MQI Scoring. The MQI requires coders who have a background in mathematics instruction, and so we contacted Harvard University for names of potential consultants. Two experienced MQI coders coded the pilot study videos during September-October 2015.

CLASS Training and Scoring. In July 2015, 15 members of the research team participated in a two-day training on the Upper Elementary Classroom Assessment Scoring System (UE CLASS). Day one of the training involved reviewing each domain in depth, discussing the observable indicators of the dimensions that comprise each domain. Next, the participants viewed a video clip and live coded according to a specific UE CLASS domain. Participants discussed their assigned score for the teacher in the clip. Each video clip had a "master score." The trainer provided the master score as well as the rationale for score. Day two

included a review of the dimensions of each domain and live coding of videos using the entire instrument. Participants discussed their assigned scores for and the trainer provided the master scores and rationale. The process continues over the course of the day to calibrate observation scores to the master coder. Within two weeks of the training, each research team member completed the online certification test. The online system included training videos to practice coding prior to taking the test. Criteria for passing the test include coding within one point of master codes on 80 percent of the codes overall, and demonstrating proficiency in each dimension by coding within one point of master codes on two out of five videos for each dimension.

Results and Discussion

Selecting an observation protocol: CLASS or MQI?

We had originally proposed using the MQI as our observation protocol, which seemed a perfect fit because of its focus on equity and math content. The most current version of the MQI no longer includes an equity scale, which decreases the alignment of the MQI with the MFA program. The MFA intervention strongly emphasizes differentiated, individualized instruction, and is less about math content than it is about enhancing teachers' awareness of and ability to tailor instruction to a student's needs. Accordingly, we looked into using the CLASS as our observation protocol.

The MFA program is designed to impact pedagogical content knowledge (PCK) and pedagogical knowledge (PK). Assessing PCK would point us to a content-specific protocol like the MQI, while assessing PK would point us to a content-neutral protocol like the CLASS.

We believe that instructional outcomes of the MFA program, with its focus on helping teachers recognize the cognitive and developmental demands of a (math-based) task and the strengths and needs of each student (Moeller & Dubitsky, 2014), can be measured using a content-neutral protocol and that instructional differences between control and treatment teachers may be addressed using the CLASS. While both the MQI and CLASS are viable options, the CLASS's emphasis on the quality of teacher-student interactions seems to be a better fit than the MQI.

We note that using the CLASS as our observation protocol does not mean we will lose the math focus. The CLASS measures instructional interactions *within the context of math instruction*. Teachers' PCK is measured by two other measures being used in the study (the Mathematical Knowledge for Teaching and the MFA Teacher Performance Assessment, respectively).

The MQI versus CLASS decision was not one that could be made solely on the basis of content versus non-content focus; we also had to consider mode, inter-rater reliability, and project budget (see the following sections).

Coding mode: Live or video?

The literature shows no clear-cut advantage in coding live or coding pre-recorded videos. For example, Casabianca et al. (2013) noted that "Both methods had large errors and low reliability… unless a large number of ratings was conducted on multiple lessons from multiple raters." On the one hand, "more ratings are needed to achieve reliable scores using video scoring than with live scoring." On the other hand, "Video observations may be more cost effective for achieving a specified level of reliability" because "additional ratings of recorded videos" is cheaper than observing more lessons.

In our experience, the pros and cons of coding live versus videos also seemed equivocal. Videotaping gives us the ability to re-watch, but live coding allows the coder to directly experience what happens during the math lesson and pick up on contextual cues. But videotaping makes teachers self-conscious and some

are fearful that the video might be used for teacher evaluations. Parents are sensitive about their children's privacy and having their children recorded, so handling opt-outs adds another layer of logistical considerations. Videotaping also requires equipment and technological savvy on the part of the classroom observer, to ensure quality of video and audio.

Achieving acceptable inter-rater reliability

The MQI offers an online training that takes upwards of 16 hours, at the end of which one takes a certification test. Math content knowledge is assessed prior to the online training. The CLASS requires coders to take a two-day, in-person training, after which an online certification test is given. We found that the MQI was very challenging, even for PhDs: a math education background really is a pre-requisite. We focus our discussion here primarily on the CLASS, but interrater reliability was a challenge even when using consultants who had achieved "calibration" with MQI master coders. We found that the two consultants did not have high levels of agreement, so we had them reconcile all the videos they coded.

Summaries of MQI and CLASS coder ratings are shown in Tables 2 and 3 (exact agreement, adjacent agreement, disagreements). Table 4 shows the results of Rasch analyses of the CLASS codes. As shown in these tables, interrater reliability was not as high as we would have desired, and varied by domain. Classroom Organization and Student Engagement, two domains that are more behavioral and arguably lower inference, showed the highest levels of agreement (particularly the Negative Climate dimension within the Classroom Organization domain). The Emotional Support and Instructional Support domains had lower levels of agreement, with combined ratings (percent exact plus percent adjacent) ranging from 66.67 to 73.08. Given these results, we decided to have each video coded by two raters, who then met to reconcile their codes.

Sensitivity of CLASS to pre-post changes. The CLASS does seem sensitive to detecting pre-post changes. We had pre-post data from 12 pilot teachers, and as shown in Table 5, there were two statistically significant differences from pre to post:

- Teachers increased in the dimension Analysis and Inquiry
- Teachers increased in the domain Instructional Support

The instructional support domain finding is of course, being driven by the analysis and inquiry dimension. But it is notable that all five of the dimensions that comprise the instructional support domain did increase in the positive direction (although analysis and inquiry was the only significant difference)

- Instructional Learning Formats: 0.06 increase, pre-post
- Content Understanding: 0.20 increase, pre-post
- Analysis and Inquiry: 0.53 increase, pre-post
- Quality of Feedback: 0.46 increase, pre-post
- Instructional Dialogue: 0.02 increase, pre-post

Making hard decisions: Optimizing study rigor within a limited budget

Although double-coding and reconciling the video data required additional staff time, we decided that the additional resources needed to have higher-quality data was called for. The fact that even experienced, MQI-calibrated coders did not have high levels of agreement with each other drove home the importance of having more than one pair of eyes rate the classroom observations. Ultimately we decided to sample and

videotape a subset of the classrooms in the main data collection, so that we could balance data quality with our limited budget.

Conclusions and Future Directions

Observing teachers' classroom practice is complex and there is no perfect observation protocol. Multiple factors must be considered, including teachers' sensitivity to being videotaped, parents' concerns about their children's privacy, and the challenge in establishing interrater reliability in coding complex classroom interactions. Although videotaping classroom observations is resource demanding, we believe that the reconciled video codes are higher quality data that will allow us to draw conclusions about the impact of MFA that are both valid and reliable.

References

Brodesky, A. R., Gross, F. E., McTigue, A. S., & Palmer, A. (2007). *A model for collaboration: Study groups are an effective way to plan math instruction for students with special needs*. Educational Leadership (Online). Alexandria, VA: ASCD.

Casabianca, J.M., McCaffrey, D.F., Gitomer, D.H., Bell, C.A., Hambre, B.K., & Pianta, R.C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757-783.

Cohen, D. K. & Hill, H. C. (1998). *Instructional policy and classroom performance: The mathematics reform in California* (CPRE RR-39). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

Corcoran, T. B. (1995). *Helping teachers teach well: Transforming professional development*. Consortium for Policy Research in Education RB-16. New Brunswick, NJ: Rutgers, State University of New Jersey.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945.

Hill. H.C. (2010). The nature and predictors of elementary teachers' Mathematical Knowledge for Teaching. *Journal for Research in Mathematics Education*, 41 (5), 513-545.

Kennedy, M. (1998). *Form and substance of in-service teacher education (Research Monograph No. 13)*. Madison, WI: National Institute for Science Education, University of Wisconsin-Madison.

MET Project. (2010). *The MQI protocol for classroom observations*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.gatesfoundation.org.

MET Project. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from www.gatesfoundation.org

Moeller, B., & Dubitsky, B. (2014). Making standards-based mathematics education accessible to students with disabilities. *Urban Perspectives*, *20*(1), 1-8.

Moeller, B., Dubitsky, B., Cohen, M., Marschke-Tobier, K., Melnick, H., & Metnetsky, L. (2012). *Mathematics for All: Facilitator Guide for Grades 3–5*. Thousand Oaks, CA: Corwin Press.

National Center for Teacher Effectiveness. (n.d.). Mathematical quality of instruction. Retrieved from http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&tabgroupid=icb.tabgroup120173.

Pianta, R. C., Hamre, B. K., & Mintz, S. L. (2012). *Classroom Assessment Scoring System---Secondary Manual.* Charlottesville, VA: Teachstone.

Yoon, K. S., Duncan, T., Lee, S. W-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007-No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs.

Table 1. Psychometric properties of CLASS and MQI

	The Classroom Assessment Scoring System™ (CLASS)	**Mathematical Quality of Instruction (MQI)**
Description of Instrument	The Upper Elementary Classroom Assessment Scoring System® (UE CLASS®) is an observational instrument developed to assess effective teacher-student interactions in grades 4-6. The CLASS dimensions are based on development theory and research suggesting that interactions between students and adults are the primary mechanism of student development and learning. The dimensions include: 1) Emotional Support; 2) Classroom Organization; 3) Instructional Support; and 4) Student Engagement. Scoring for the CLASS dimensions is not determined by the presence of materials, the physical environment or safety, or the adoption of a specific curriculum. In the Upper Elementary CLASS, the focus centers on what teachers *do* with the materials they have as well as the interactions that teachers have with their students.	The Mathematical Quality of Instruction (MQI) is designed to reliably measure the mathematical work that occurs in classrooms, on the theory that that work is distinct from classroom climate, pedagogical style, or the deployment of generic instructional strategies. The MQI is based on a theory of instruction that focuses on resources and their use, existing literature on effective instruction in mathematics, and on an analysis of nearly 250 videotapes of diverse teachers and teaching. The MQI measures the mathematical quality of instruction by assessing the relationship among the teacher, the student, and mathematical content using five elements: 1) richness of the mathematics; 2) errors and imprecision; 3) working with students and mathematics; 4) student participation in meaning-making and reasoning; and 5) connections between classroom work and mathematics. Each element is used to help assess one of three relationships: teacher-content, teacher-student, or student-content.
Observation Process	Observers typically watch a lesson for 15 minutes, taking notes on the specific behaviors they observe related to each of the CLASS dimensions. Scoring is completed at the dimension level using a 7-point scale, with the low range being a score of 1-2, the middle range 3-5, and the high range 6-7. The CLASS manual provides detailed information to help observers determine the specific score. The observer then watches the next 15 minutes and scores each of the dimensions again, repeating this cycle of observation and scoring until the end of the lesson. Lesson scores are created by averaging scores across all 15-minute cycles, and scores for teachers are averaged across lessons. Observations can be scored live or using video.	The MQI protocol is designed primarily for use assessing videotaped instruction. Each videotaped lesson is divided into roughly equal-length five- to seven-and-a-half-minute segments for scoring. Raters assign each segment a score for each of the five MQI elements, and assign the whole lesson an overall MQI score. Two raters working independently of one another score each lesson, and scores are averaged across lessons to comprise a teacher score.
Validity Evidence	The CLASS was developed based on extensive research on classroom practices shown to relate to students' social and academic development in schools. The dimensions were derived from a review of constructs assessed in classroom observation instruments used in school research, literature on effective teaching practices, focus groups, and extensive piloting. To test the degree to which data from actual classrooms matched the theoretical framework, confirmatory factor analyses were conducted across three studies, consisting of 1,493 classrooms across multiple states. The factor loadings were in the moderate to high range (.73 or higher). In addition, numerous experts in classroom quality and teaching effectiveness have agreed that the CLASS tool measures aspects of the classroom that are essential in determining student performance, suggesting adequate face validity. The CLASS has been used to observe over 20,000 classrooms across the United States. To assess criterion validity, the relationship between the	The MQI instrument is designed to provide information about the quality of teachers' enactment of mathematics instruction. For all dimensions except Errors, higher scores indicate better performance; for Errors, higher scores indicate more problematic instruction. *Construct Validity*: Factor analyses supported theoretical constructs. *Criterion Validity*: MQI scores were significantly related to teacher MKT scores (Responds to Students r = .65; Errors r = -.83). This is not surprising as the MQI was initially designed to validate the Mathematical Knowledge for Teaching (MKT). The MQI was also correlated with other observation measures: • UTEACH Teacher Observation Protocol (UTOP); r = .85 - a measure of math specific teaching • Framework for Teaching (FFT); r = .67 • Classroom Assessment Scoring System (CLASS); r = .69

	The Classroom Assessment Scoring System™ (CLASS)	Mathematical Quality of Instruction (MQI)
	Upper Elementary and Secondary CLASS and various other measures of classrooms: • Framework for Teaching (FFT); *r* = .88 • Mathematical Quality of Instruction (MQI); *r* = .69 • UTEACH Teacher Observation Protocol (UTOP); *r* = .68 • Protocol for Language Arts Teaching Observations (PLATO); *r* = .86 The CLASS measure was designed to assess classroom-level processes that are directly associated with students' performance. Several studies provide evidence of predictive validity; the teachers who demonstrated the types of practices emphasized in the CLASS measure had higher value-added scores than teachers who did not.	This tool was validated by randomly assigning teachers to classrooms, collecting data using multiple measures, and testing whether MQI scores predicted student outcomes. Several studies indicate that MQI scores are significantly related to teacher value-added scores.
Reliability Evidence	Evidence suggests that CLASS scores, assigned by trained, certified observers, are reliable. Three studies observed rater agreement on three dimensions: Emotional Support (percent agreement ranged from 77 to 89%), Classroom Organization (percent agreement ranged from 83 to 86%), and Instructional Support (percent agreement ranged from 73 to 75%). The internal consistency estimates for the CLASS domains indicate that the dimensions comprising each domain tap into consistent characteristics of classrooms. When measured in the fall and spring, CLASS scores have low to moderate correlations, indicating moderate stability over time. Finally, when two observers code the same cycle, they consistently assign scores that are within one point on the scale (an exact match 30% of the time, agreement within one point on the scale ranges from 64% to 98%). To become certified, observers attend a two-day CLASS Observation Training. During this training, observers learn about CLASS domains and dimensions, then watch and code multiple, videotaped lesson segments that have been master-coded by a team of CLASS experts. Over the course of the two days, trainees calibrate their scoring to be in line with the master coders' scores. After the training, potential users take a reliability test, which involves independently watching and coding an additional five videotaped lesson segments. Criteria for passing the test include coding within one point of master codes on 80% of the codes overall, and demonstrating proficiency in each dimension by coding within one point of master codes on two out of five videos for each dimension.	Teacher-level reliability, 3 lessons 2 raters: Richness (.80), Errors (.75), Working with students (.68), Student participation (.82), and Overall composite score (.77)
References	• Hamre, B. K. (2011). *Using Classroom Observation to Gauge Teacher Effectiveness: Classroom Assessment Scoring System (CLASS)*. Presentation for the National Center for Teacher Effectiveness. Cambridge, MA. http://www.gse.harvard.edu/cepr-resources/files/news-events/ncte-conference-class-hamre.pdf • Kane, T.J., & Staiger, D.O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement*	• Hill, H. C. (2011). *Mathematical Quality of Instruction (MQI).* Presentation for the National Center for Teacher Effectiveness. Cambridge, MA. http://www.gse.harvard.edu/cepr-resources/files/news-events/ncte-conference-mqi-hill.pdf • Hill, H.C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G.C., Sleep, L., & Ball, D.L. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An exploratory

	The Classroom Assessment Scoring System™ (CLASS)	Mathematical Quality of Instruction (MQI)
	Gains. Seattle, WA: Bill & Melinda Gates Foundation. http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf • MET Project. (2010). *The CLASS protocol for classroom observations.* Seattle, WA: Bill and Melinda Gates Foundation. http://metproject.org/resources/CLASS_10_29_10.pdf • Pianta, R. C., Hamre, B. K., & Mintz, S. L. (2012). *Classroom Assessment Scoring System – Upper Elementary Manual*. Charlottesville, VA: Teachstone. • Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109-119.	study. *Cognition and Instruction*, *26*, 430-511. • Kane, T.J., & Staiger, D.O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Seattle, WA: Bill & Melinda Gates Foundation. http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf • Learning Mathematics for Teaching Project. (2011). Measuring the Mathematical Quality of Instruction. *Journal of Mathematics Teacher Education, 14*(1), 25-47. • MET Project. (2010). *The MQI protocol for classroom observations.* Seattle, WA: Bill and Melinda Gates Foundation. http://metproject.org/resources/MQI_10_29_10.pdf

Table 2. MQI inter-rater reliability

	Total Exact	Total Adjacent	Total Exact + Adjacent	Total Disagree	Percent Exact	Percent Adjacent	Percent Exact + Adjacent	Percent Disagree
Richness of Mathematics								
Linking Between Representations	79	29	108	4	70.54	25.89	96.43	3.57
Explanations	62	37	99	13	55.36	33.04	88.39	11.61
Mathematical Sense-Making	58	46	104	8	51.79	41.07	92.86	7.14
Multiple Procedures or Solution Methods	82	27	109	3	73.21	24.11	97.32	2.68
Patterns and Generalizations	110	2	112	0	98.21	1.79	100.00	0.00
Mathematical Language	51	60	111	1	45.54	53.57	99.11	0.89
Overall Richness of the Mathematics	71	39	110	2	63.39	34.82	98.21	1.79
Working with Students and Mathematics								
Remediation of Student Errors and Difficulties	50	55	105	7	44.64	49.11	93.75	6.25
Teacher Uses Student Mathematical Contributions	33	75	108	4	29.46	66.96	96.43	3.57
Overall Working with Students and Mathematics	48	60	108	4	42.86	53.57	96.43	3.57
Errors and Imprecision								
Mathematical Content Errors	104	6	110	2	92.86	5.36	98.21	1.79
Imprecision in Language or Notation	98	13	111	1	87.50	11.61	99.11	0.89
Lack of Clarity in Presentation of Mathematical Content	99	11	110	2	88.39	9.82	98.21	1.79
Overall Errors and Imprecision	84	26	110	2	75.00	23.21	98.21	1.79
Common Core Aligned Student Practices								
Students Provide Explanations	58	48	106	6	51.79	42.86	94.64	5.36
Student Mathematical Questioning and Reasoning (SMQR)	55	50	105	7	49.11	44.64	93.75	6.25
Students Communicate about the Mathematics of the Segment	43	69	112	0	38.39	61.61	100.00	0.00
Task Cognitive Demand	78	29	107	5	69.64	25.89	95.54	4.46
Students Work with Contextualized Problems	83	28	111	1	74.11	25.00	99.11	0.89
Overall Common Core Aligned Student Practices	66	43	109	3	58.93	38.39	97.32	2.68
Whole Lesson Codes								
Lesson Time is Used Efficiently	40	49	89	23	35.71	43.75	79.46	20.54
Lesson is Mathematically Dense	95	12	107	5	84.82	10.71	95.54	4.46
Students are Engaged	55	56	111	1	49.11	50.00	99.11	0.89
Lesson Contains Rich Mathematics	45	52	97	15	40.18	46.43	86.61	13.39
Teacher Attends to and Remediates Student Difficulty	40	50	90	22	35.71	44.64	80.36	19.64
Teacher Uses Student Ideas	12	58	70	42	10.71	51.79	62.50	37.50
Mathematics is Clear and not Distorted	81	21	102	10	72.32	18.75	91.07	8.93
Tasks and Activities Develop Mathematics	89	22	111	1	79.46	19.64	99.11	0.89
Lesson Contains Common Core Aligned Student Practices	45	56	101	11	40.18	50.00	90.18	9.82
Overall MQI								
	80	32	112	0	71.43	28.57	100.00	0.00

Table 3. CLASS inter-rater reliability

	Total Exact	Total Adjacent	Total Exact + Adjacent	Total Disagree	Percent Exact	Percent Adjacent	Percent Exact + Adjacent	Percent Disagree
Emotional Support								
Positive Climate	28	25	53	25	35.90	32.05	67.95	32.05
Teacher Sensitivity	20	32	52	26	25.64	41.03	66.67	33.33
Regard for Student Perspectives	20	34	54	24	25.64	43.59	69.23	30.77
Classroom Organization								
Behavior Management	43	22	65	13	55.13	28.21	83.33	16.67
Productivity	37	29	66	12	47.44	37.18	84.62	15.38
Negative Climate	59	18	77	1	75.64	23.08	98.72	1.28
Instructional Support								
Instructional Learning Formats	29	27	56	22	37.18	34.62	71.79	28.21
Content Understanding	23	30	53	25	29.49	38.46	67.95	32.05
Analysis and Inquiry	25	32	57	21	32.05	41.03	73.08	26.92
Quality of Feedback	16	36	52	26	20.51	46.15	66.67	33.33
Instructional Dialogue	25	27	52	26	32.05	34.62	66.67	33.33
Student Engagement								
Active Engagement	18	45	63	15	23.08	57.69	80.77	19.23

Table 4. Results of Rasch analyses of CLASS coder ratings

Emotional Support

Exact agreement was poor (31.2%) but higher than expected exact agreement. Rater 7 tended to rate observations higher than the group of raters.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
394	81	4.86	4.84	1.12	.12	.71	-2.0	.73	-1.9	1.27	.74	.74	32.1	29.2	7	7
464	99	4.69	4.81	1.09	.11	1.30	2.0	1.30	2.0	.66	.76	.76	29.3	29.5	9	9
506	108	4.69	4.76	1.04	.10	1.06	.4	1.04	.3	1.02	.75	.77	28.7	30.0	4	4
223	51	4.37	4.68	.96	.15	1.18	.9	1.21	1.0	.68	.63	.73	39.2	29.5	6	6
84	18	4.67	4.43	.72	.25	1.19	.6	1.07	.3	.90	.71	.70	50.0	29.4	10	10
42	9	4.67	4.40	.70	.35	.18	-2.7	.17	-2.8	1.99	.93	.71	44.4	29.1	8	8
387	84	4.61	4.21	.52	.12	1.36	2.2	1.32	1.9	.56	.71	.75	28.6	30.1	1	1
372	90	4.13	4.20	.51	.11	.88	-.8	.87	-.8	1.20	.81	.77	37.8	30.3	2	2
277	66	4.20	4.09	.40	.13	.74	-1.5	.75	-1.5	1.24	.82	.79	18.2	30.2	5	5
247	63	3.92	4.00	.32	.13	.54	-3.2	.59	-2.7	1.39	.68	.68	31.7	27.7	3	3
299.6	66.9	4.48	4.44	.74	.16	.92	-.4	.90	-.4		.75				Mean (Count: 10)	
146.1	31.2	.29	.30	.28	.07	.36	1.9	.34	1.7		.08				S.D. (Population)	
154.0	32.9	.31	.31	.30	.08	.38	2.0	.36	1.8		.08				S.D. (Sample)	

Model, Populn: RMSE .17 Adj (True) S.D. .22 Separation 1.28 Strata 2.04 Reliability (not inter-rater) .62
Model, Sample: RMSE .17 Adj (True) S.D. .24 Separation 1.39 Strata 2.19 Reliability (not inter-rater) .66
Model, Fixed (all same) chi-square: 54.2 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-square: 7.8 d.f.: 8 significance (probability): .46
Inter-Rater agreement opportunities: 333 Exact agreements: 104 = 31.2% Expected: 98.6 = 29.6%

It was easiest to rate *teacher sensitivity* (average logit = -.88) but more difficult to rate *regard for student perspectives* (average logit = .92).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Dimension
816	223	3.66	3.52	.92	.07	1.08	.8	1.10	1.0	.92	.62	.67	3	Regard for Student Perspectives
1009	223	4.52	4.49	-.04	.07	.91	-1.0	.90	-1.1	1.10	.70	.70	1	Positive Climate
1171	223	5.25	5.32	-.88	.07	.98	-.1	.96	-.3	1.00	.72	.69	2	Teacher Sensitivity
998.7	223.0	4.48	4.44	.00	.07	.99	-.1	.98	-.2		.68		Mean (Count: 3)	
145.1	.0	.65	.74	.73	.00	.07	.8	.08	.9		.04		S.D. (Population)	
177.7	.0	.80	.90	.90	.00	.09	.9	.10	1.1		.05		S.D. (Sample)	

Model, Populn: RMSE .07 Adj (True) S.D. .73 Separation 10.08 Strata 13.78 Reliability .99
Model, Sample: RMSE .07 Adj (True) S.D. .89 Separation 12.37 Strata 16.83 Reliability .99
Model, Fixed (all same) chi-square: 299.9 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

Classroom Organization

This domain had the highest exact agreement (61.6%). Rater 3 tended to rate observations higher than the group of raters and would be considered the most lenient.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
426	63	6.76	6.80	2.80	.29	1.12	.4	1.01	.1	.81	.46	.50	66.7	66.8	3	3
586	90	6.51	6.63	2.13	.17	.90	-.4	.89	-.3	1.06	.57	.55	52.2	55.6	2	2
529	81	6.53	6.60	2.05	.19	.87	-.5	.73	-.7	1.13	.66	.63	64.1	65.6	7	7
420	66	6.36	6.56	1.93	.19	.54	-2.4	.46	-2.2	1.33	.75	.64	69.7	55.9	5	5
111	18	6.17	6.49	1.76	.32	1.35	.9	2.18	2.1	.16	.42	.63	55.6	50.6	10	10
648	99	6.55	6.48	1.73	.17	1.30	1.3	1.06	.2	.89	.58	.60	64.6	67.2	9	9
536	84	6.38	6.48	1.73	.17	1.31	1.4	1.36	1.3	.82	.59	.64	60.7	60.6	1	1
62	9	6.89	6.34	1.45	1.07	.72	.0	.57	.0	1.10	.49	.41	100.0	86.7	8	8
667	108	6.18	6.30	1.39	.13	.77	-1.4	.64	-2.0	1.25	.72	.64	56.5	52.0	4	4
316	51	6.20	6.21	1.23	.20	1.43	1.6	1.25	.8	.74	.60	.66	58.8	55.1	6	6
430.1	66.9	6.45	6.49	1.82	.29	1.03	.1	1.01	.0		.58				Mean (Count: 10)	
199.8	31.2	.23	.17	.43	.26	.29	1.3	.48	1.3		.10				S.D. (Population)	
210.7	32.9	.25	.17	.45	.28	.31	1.3	.50	1.4		.11				S.D. (Sample)	

Model, Populn: RMSE .39 Adj (True) S.D. .17 Separation .43 Strata .91 Reliability (not inter-rater) .16
Model, Sample: RMSE .39 Adj (True) S.D. .22 Separation .57 Strata 1.09 Reliability (not inter-rater) .24
Model, Fixed (all same) chi-square: 36.6 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-square: 7.2 d.f.: 8 significance (probability): .51
Inter-Rater agreement opportunities: 333 Exact agreements: 205 = 61.6% Expected: 199.4 = 59.9%

It was easiest to rate *negative climate* (average logit = -1.13) but more difficult to rate *productivity* (average logit = .71).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Dimension
1377	223	6.17	6.14	.71	.09	1.03	.2	.98	-.1	.87	.65	.66	2	Productivity
1408	223	6.31	6.31	.43	.10	.97	-.2	.93	-.5	1.11	.66	.62	1	Behavior Management
1516	223	6.80	6.83	-1.13	.16	1.10	.6	.88	-.4	1.04	.43	.41	3	Negative Climate
1433.7	223.0	6.43	6.43	.00	.12	1.03	.2	.93	-.4		.58		Mean (Count: 3)	
59.6	.0	.27	.29	.81	.03	.05	.3	.04	.2		.11		S.D. (Population)	
73.0	.0	.33	.36	.99	.04	.06	.4	.05	.2		.13		S.D. (Sample)	

Model, Populn: RMSE .12 Adj (True) S.D. .80 Separation 6.66 Strata 9.21 Reliability .98
Model, Sample: RMSE .12 Adj (True) S.D. .98 Separation 8.19 Strata 11.25 Reliability .99
Model, Fixed (all same) chi-square: 102.0 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

Instructional Support

Exact agreement was poor (32.8%) but higher than expected exact agreement. Rater 7 tended to rate observations higher than the group of raters.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
561	135	4.16	4.02	-.32	.09	.83	-1.4	.86	-1.1	1.08	.69	.74	30.8	28.5	7	7
668	180	3.71	3.92	-.41	.08	.89	-1.1	.90	-1.0	1.19	.79	.76	34.4	29.2	4	4
304	85	3.58	3.89	-.44	.11	.93	-.4	.94	-.3	1.10	.71	.69	41.2	28.0	6	6
440	110	4.00	3.87	-.45	.10	.78	-1.7	.77	-1.8	1.25	.82	.76	31.8	31.0	5	5
109	30	3.63	3.35	-.91	.19	.56	-2.0	.60	-1.8	1.41	.85	.72	36.7	28.7	10	10
345	105	3.29	3.34	-.92	.10	.49	-4.8	.50	-4.6	1.58	.73	.62	31.4	27.2	3	3
477	140	3.41	3.34	-.92	.09	1.41	3.1	1.40	3.1	.51	.65	.72	32.1	28.7	1	1
473	150	3.15	3.20	-1.05	.09	.90	-.8	.89	-.9	1.15	.76	.74	37.3	29.5	2	2
502	165	3.04	2.87	-1.36	.08	1.60	4.8	1.56	4.5	.31	.60	.68	27.3	26.8	9	9
45	15	3.00	2.45	-1.82	.26	.55	-1.4	.55	-1.4	1.54	.70	.61	13.3	23.1	8	8
392.4	111.5	3.50	3.43	-.86	.12	.89	-.6	.90	-.6		.73				Mean (Count: 10)	
185.3	52.1	.37	.48	.46	.06	.34	2.6	.33	2.5		.07				S.D. (Population)	
195.3	54.9	.39	.51	.48	.06	.36	2.7	.35	2.6		.08				S.D. (Sample)	

Model, Populn: RMSE .13 Adj (True) S.D. .44 Separation 3.31 Strata 4.74 Reliability (not inter-rater) .92
Model, Sample: RMSE .13 Adj (True) S.D. .46 Separation 3.50 Strata 5.00 Reliability (not inter-rater) .92
Model, Fixed (all same) chi-square: 149.0 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-square: 7.9 d.f.: 8 significance (probability): .45
Inter-Rater agreement opportunities: 555 Exact agreements: 182 = 32.8% Expected: 158.3 = 28.5%

It was easiest to rate *instructional learning formats* (average logit = -.87) but more difficult to rate *analysis and inquiry* (average logit = 1.02).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Dimension
578	223	2.59	2.40	1.02	.07	1.03	.3	1.05	.5	1.00	.71	.69	3	Analysis and Inquiry
736	223	3.30	3.17	.22	.07	1.01	.1	.97	-.3	1.04	.73	.70	5	Instructional Dialogue
820	223	3.68	3.61	-.18	.07	1.19	2.0	1.19	2.0	.75	.66	.69	4	Quality of Feedback
824	223	3.70	3.63	-.19	.07	.97	-.3	.99	-.1	1.00	.65	.69	2	Content Understanding
966	223	4.33	4.37	-.87	.07	.77	-2.6	.78	-2.5	1.22	.68	.66	1	Instructional Learning Formats
784.8	223.0	3.52	3.43	.00	.07	.99	-.1	.99	-.1		.69		Mean (Count: 5)	
127.1	.0	.57	.65	.62	.00	.13	1.5	.13	1.5		.03		S.D. (Population)	
142.1	.0	.64	.72	.69	.00	.15	1.7	.15	1.7		.03		S.D. (Sample)	

Model, Populn: RMSE .07 Adj (True) S.D. .62 Separation 8.77 Strata 12.03 Reliability .99
Model, Sample: RMSE .07 Adj (True) S.D. .69 Separation 9.82 Strata 13.43 Reliability .99
Model, Fixed (all same) chi-square: 365.1 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-square: 4.0 d.f.: 3 significance (probability): .27

Student Engagement

Exact agreement was poor (28.8%) and lower than expected exact agreement. Rater 10 tended to rate observations higher than the group of raters.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
31	6	5.17	5.79	2.19	.81	1.00	.2	.89	.0	1.11	.95	.92	16.7	42.2	10	10
114	22	5.18	5.19	.38	.38	.82	-.5	.78	-.7	1.31	.87	.86	22.7	45.9	5	5
88	17	5.18	5.16	.30	.43	.99	.0	.97	.0	.97	.89	.86	29.4	46.1	6	6
142	28	5.07	5.13	.21	.33	1.45	1.6	1.38	1.3	.60	.75	.82	28.6	46.7	1	1
177	36	4.92	4.95	-.25	.29	1.15	.6	1.17	.7	.77	.80	.80	33.3	48.9	4	4
98	21	4.67	4.77	-.72	.37	.68	-1.1	.72	-.9	1.28	.73	.75	33.3	46.0	3	3
135	27	5.00	4.76	-.74	.33	.73	-1.0	.74	-1.0	1.26	.73	.77	15.4	44.3	7	7
156	33	4.73	4.74	-.80	.31	.76	-1.0	.76	-1.0	1.25	.80	.80	33.3	47.5	9	9
138	30	4.60	4.73	-.83	.33	1.32	1.1	1.25	.9	.77	.80	.81	30.0	50.4	2	2
14	3	4.67	4.49	-1.47	.95	.20	-1.3	.21	-1.3	1.91	.98	.51	66.7	42.0	8	8
109.3	22.3	4.92	4.97	-.17	.45	.91	-.2	.88	-.2		.83				Mean (Count: 10)	
50.1	10.4	.22	.35	.97	.22	.34	1.0	.31	.9		.09				S.D. (Population)	
52.8	11.0	.23	.37	1.02	.23	.36	1.1	.33	1.0		.09				S.D. (Sample)	

Model, Populn: RMSE .51 Adj (True) S.D. .83 Separation 1.63 Strata 2.51 Reliability (not inter-rater) .73
Model, Sample: RMSE .51 Adj (True) S.D. .89 Separation 1.76 Strata 2.67 Reliability (not inter-rater) .75
Model, Fixed (all same) chi-square: 26.4 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-square: 6.2 d.f.: 8 significance (probability): .62
Inter-Rater agreement opportunities: 111 Exact agreements: 32 = 28.8% Expected: 52.2 = 47.0%

Table 5. CLASS ratings, pretest to posttest for MFA pilot (n=12)

Domain/Dimension	Pretest Mean (SE)	Posttest Mean (SE)	Difference (SE)	t-value (df)
Emotional Support	4.54 (0.17)	4.37 (0.17)	-0.18 (0.17)	-1.07 (11)
Positive Climate Relationships; positive affect; positive communications; respect	4.66 (0.20)	4.30 (0.20)	-0.35 (0.17)	-2.02 (11)
Teacher Sensitivity Awareness; responsiveness to academic and social/emotional needs and cues; effectiveness in addressing problems; student comfort	5.24 (0.22)	5.25 (0.22)	0.01 (0.27)	0.04 (11)
Regard for Student Perspectives Flexibility and student focus; connections to real life; support for autonomy and leadership; meaningful peer interactions	3.74 (0.19)	3.54 (0.19)	-0.19 (0.24)	-0.80 (11)
Classroom Organization	6.46 (0.11)	6.37 (0.11)	-0.09 (0.11)	-0.79 (11)
Behavior Management Clear expectations; proactive; effective redirection of misbehavior; student behavior	6.34 (0.20)	6.21 (0.20)	-0.13 (0.15)	-0.84 (11)
Productivity Maximizing learning time; routines; transitions; preparation	6.18 (0.15)	6.16 (0.15)	-0.02 (0.13)	-0.13 (11)
Positive Climate (Negative Climate, reverse scored) Absence of negative affect; punitive control; disrespect	6.86 (0.08)	6.75 (0.08)	-0.11 (0.10)	-1.09 (11)
Instructional Support	3.38 (0.14)	3.63 (0.14)	0.25 (0.10)	2.45* (11)
Instructional Learning Formats Learning targets/organization; variety of modalities, strategies, and materials; active facilitation; effective engagement	4.31 (0.17)	4.37 (0.17)	0.06 (0.22)	0.27 (11)
Content Understanding Depth of understanding; communication of concepts and procedures; background knowledge and misconceptions; transmission of content knowledge and procedures; opportunity for practice of procedures and skills	3.62 (0.15)	3.82 (0.15)	0.20 (0.12)	1.67 (11)
Analysis and Inquiry Facilitation of higher-order thinking; opportunities for novel application; metacognition	2.34 (0.18)	2.86 (0.18)	0.53 (0.16)	3.32* (11)
Quality of Feedback Feedback loops; scaffolding; building on student responses; encouragement and affirmation	3.44 (0.21)	3.90 (0.21)	0.46 (0.24)	1.92 (11)
Instructional Dialogue Cumulative content-driven exchanges; distributed talk; facilitation strategies	3.19 (0.21)	3.21 (0.21)	0.02 (0.16)	0.11 (11)
Student Engagement	4.91 (0.16)	4.86 (0.16)	-0.05 (0.16)	-0.31 (11)