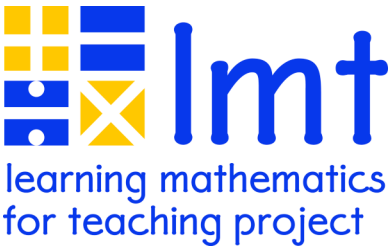
A Coding Rubric for Measuring the Mathematics Quality of Instruction

Learning Mathematics for Teaching Technical Report #LMT1.06

University of Michigan

DRAFT



Learning Mathematics for Teaching
University of Michigan
School of Education
610 E. University #1600
Ann Arbor, MI 48109-1259
www.sitemaker.umich.edu/lmt

Citation:

Learning Mathematics for Teaching (2006). *A Coding rubric for Measuring the Mathematical Quality of Instruction* (Technical Report LMT1.06*)*. Ann Arbor, MI: University of Michigan, School of Education.

Research in this paper was supported by NSF grants REC-0207649, EHR-0233456, and EHR-0335411. The Learning Mathematics Teaching project consists of: Heather C. Hill, Deborah Loewenberg Ball, Hyman Bass & Stephen Schilling, Principal Investigators; Merrie Blunk, Catherine Brach, Charalambos Charalambous, Carolyn Dean, Seán Delaney, Jennifer Lewis, Imani Masters-Goffney, Geoffrey Phelps, Laurie Sleep, Mark Thames, and Deborah Zopf, research staff. We thank Hilda Borko, Paul Cobb, and Nicole Kersting for reading and helping improve a previous draft.

What mathematical knowledge do teachers need to successfully work with students, and how do we know when they have it? Over the past several years, the first part of this question has been the subject of extensive reports, research, policy initiatives and, often, debate. The second part of this question has also seen increasing treatment by the testing and scholarly community, as researchers have developed open-ended, interview, and multiple-choice assessments of mathematical knowledge for teaching. However, none of these methods is satisfactory in one critical way, in that none can actually measure the *quality of the mathematics in actual classroom instruction*. Teachers' performance on pencil-and-paper assessments (or oral interview tasks) may or may not correlate with what they can actually do with real-life content, materials, and students. Yet observational studies, which have historically been used to study the nature of mathematical knowledge used in classrooms, have not to date been designed to provide reliable estimates for large numbers of individual teachers.

Because of a growing need among researchers to study changes in the mathematical quality of teachers' practice, we argue for an observation-based instrument that can quantify the *quality of the mathematics in instruction*. By quality of mathematics in instruction, we mean the extent of key mathematical characteristics in a lesson, including accurate use of mathematical language, the avoidance of mathematical errors or oversights, the provision of mathematical explanations when warranted, the connection of classroom work to important mathematical ideas, and the work of ensuring all students access to mathematics. We argue that this constellation of lesson characteristics is one key– and by far the most important– indicator of an individual's mathematical knowledge for teaching. We also argue this instrument needs to be largely distinct from measures of teachers' pedagogical choices, a frequent focus of observational rubrics; rather than measuring alignment with "reform" teaching, in other words, a measure is needed that quantifies the quality of the mathematics appearing in instruction, whatever the teaching method.

In this paper, we describe our efforts to develop such a rubric. These efforts began as an attempt to validate pencil-and-paper measures of teachers' mathematical knowledge for teaching (MKT) (LMT, 2006); teachers who completed our pencil-and-paper MKT measures would be videotaped, those videotaped lessons "graded," and the scores on both instruments correlated. To "grade" the videotapes, however, we needed a detailed rubric mapping how mathematical knowledge for teaching might appear in practice, as teachers worked with curriculum and students. This effort quickly became a measures development project in its own right, as well as fertile ground for exploring and naming new elements of MKT.

This technical report is intended mainly for potential users of the instrument, to introduce this audience to our theoretical foundation, codes, and procedures. We begin with a review of past and contemporary uses of observation to understand teachers' knowledge, including a description of our needs in this arena. We discuss our specific codes and coding protocol, the development of our instrument, and suggest some directions for analysis. We review our early findings, then consider issues related to the adoption and use of this instrument in different locations.

Assessing Teachers' Mathematical Knowledge via Observation: A Brief Overview

Observation of classroom teaching has been fruitfully used for over two decades to explore the *territory* of mathematical knowledge for teaching. Beginning in 1985, when Lee Shulman and

colleagues proposed that teachers need not only knowledge of content itself but also pedagogical content knowledge, scholars and other observers have been engaged in using observational techniques to uncover aspects of that knowledge.

Early observational research shared some common elements. Researchers typically collected tens, if not hundreds, of hours of observations or videotapes; however, most published research typically focused on a tiny fraction of the data, often even just a few minutes. Analysis was primarily qualitative, as opposed to quantitative, with researchers using methods and coding systems tailored specifically to the mathematical topics and questions at hand. And scholars typically combined observational records with other sources of data to gain insight into mathematical knowledge for teaching.

One classic example of research in this tradition is Leinhardt and Smith's (1985) study of expertise in mathematics instruction. To explore the relationship between teacher behavior and subject matter knowledge, the authors studied eight teachers intensively, collecting three months of observational field notes from their mathematics lessons, 10 hours of videotaped lessons, interviews on the videotaped lesson and other topics, and teacher performance on a card sort task. They used the observations of instruction to construct a measure, ranking teachers' knowledge as high, medium, or low based on "in-class discussions over 3 years and by considering their presentations and explanations as well as their errors" (p. 251). This strategy—sorting teachers by their actual in-class mathematical performance—is rare in the literature, and, unfortunately, not well explicated in their published work. The authors then examined teachers' knowledge in light of performance on interview tasks, and examined three teachers' teaching of fractions in much more depth by intensive description of single lessons on simplifying fractions. This method—thick description—would prove a mainstay in exploring mathematical knowledge for teaching.

Borko, Eisenhart et al. (1992) provides another example of the observational method for measuring teachers' knowledge of mathematics. In it, the authors focused on a few minutes in the course of an hour-long review lesson, moments in which a student teacher was asked by a student to explain the division of fractions algorithm. Audiotapes of the lesson enhanced field notes taken by live observers, and allowed the construction of a dependent variable of sorts: an assessment of this teacher's capacity to provide a conceptually based justification for the standard algorithm. The authors noted that, in practice, this teacher's mathematical knowledge of division of fractions was poor; she used a concrete model when in fact such models are not in general good models for explaining why this particular algorithm works, and her model represented *multiplication,* not division. The authors then used data from interviews, her performance on open-ended mathematics problems, and records from this teacher's preservice education program to explain her in-class performance.

Sherin (1996), Thompson & Thompson (1994), and others have conducted similar studies. An important feature of much of this early work was to explore the *nature* of teachers' mathematical knowledge. However, as scholars and policy-makers began to suspect that not all teachers had strong mathematical knowledge, interest increased in studies of how teachers might learn this content area, and how such knowledge related to other characteristics such as certification, coursework, and student achievement. Some studies required measures of teachers' mathematical knowledge that could be used at scale—that is, instruments that could be used across hundreds or thousands of teachers at multiple time points. Many of these studies have designed multiple-choice instruments for this use. However, other studies working with smaller samples of teachers could clearly benefit from an instrument with more face validity, but one

that also returns reliable measures of knowledge for multiple teachers, occasions, and data collection sites.

Our project, *Learning Mathematics for Teaching,* is one such study. We investigate the mathematical knowledge needed for teaching, and how such knowledge develops as a result of experience and professional learning. As part of that work, we write, pilot, and analyze multiple choice items that reflect real mathematics tasks teachers face in classrooms. These measures are different from conventional mathematics tests in that they not only assess whether teachers can solve the problems they directly teach children, but also how they work through some of the mathematical tasks unique to teaching – for instance, assessing student work, representing numbers and operations, and explaining common mathematical rules or procedures (Ball & Bass, 2003; Ball, Hill & Bass, 2005). Assessments composed of such tasks are used to measure the effectiveness of professional development intended to improve teachers' mathematical knowledge.

However, few would agree that teachers' performance on a pencil-and-paper mathematics assessment necessarily predicts their in-class performance. Many have critiqued multiple-choice measures such as ours on the premise that no test cast in a multiple choice format could measure a complex and judgment-laden practice such as teaching (Berliner, 2005). In a 2002 issue of *English Education*, nearly all articles railed against the very idea of standardized testing for teachers:

> Virtually all of the criticisms leveled against testing in schools also apply to the quick and dirty attempt to demand accountability in testing teachers. Timed tests given to children are really evaluating speed rather than thoughtfulness, and the same is true when they're given to adults. Multiple choice tests and contrived open response items are not meaningful ways of assessing how much students understand, and neither are they particularly effective in telling us how well educators can educate. (Kohn, quoted in Appelman and Thompson, 2002, p. 96.)

To be sure, very little work has been done examining the predictive validity of many current teacher tests—or said another way, how well tests like the Praxis or NES predict how well a teacher will teach, or how much growth she will foster in her students. Because we found ourselves interested in whether and how teachers' performance on our own multiple-choice assessment related to these things, we designed our validation work around these questions. We found that student performance was related to MKT as assessed by our pencil-and-paper measures (Hill, Rowan, & Ball, 2005). Next, we began a study to understand how teachers' pencil-and-paper assessment performance would relate to classroom performance. Our initial goal was to "correlate" teachers' pencil-and-paper scores with the mathematical aspects of their actual classroom teaching in order validate our multiple choice measure of MKT.

Our literature review revealed that in the past several years, several standardized protocols for observing the characteristics of mathematics instruction (or videotaped records of instruction) have emerged. Leaving aside instruments that focus solely on the pedagogical aspects of teaching mathematics—i.e., the degree to which students work in groups, work on extended investigations, or answer questions—two seemed plausible for our use: the *Reformed Teaching Observation Protocol* (RTOP; Sawada & Pilburn, 2000) and *Inside the Classroom Observation and Analytic Protocol* (Horizon Research, 2000).[1] Both instruments ask for ratings of the extent to which content is presented accurately. Both ask whether the content presented to students is

[1] INTASC and NBPTS also have elementary-level rubrics for scoring teacher portfolios entries (which include video). See Porter, Youngs & Odden (2001) for more information.

mathematically interesting and worthwhile. And both ask about some elements of what some consider "rich" instruction—the use of representations, explanations, and abstractions, for instance. As such, these instruments contain elements that some might employ to measure the *quality of the mathematics in instruction*, or the accuracy of content, richness of representation and explanation, and connectedness of classroom tasks to mathematical principles. From this measure of knowledge use in particular lessons, one might infer the teacher's grasp of mathematical knowledge for teaching.

RTOP and Horizon's instruments, however, both embed the ratings of teacher knowledge in larger scales intended to measure the extent to which classroom instruction aligns with the National Council for Teachers of Mathematics standards. These two instruments are designed, as per their materials, to measure the *quality of mathematics instruction*, including the richness and correctness of mathematical content *and* the way material is conveyed to students (e.g., presence of collaborative learning approaches, investigations, higher order questions, student reflection). In these packages, no direct measure of teacher's mathematical knowledge in practice is available; instead, teacher knowledge is estimated as a component of how mathematical material is presented to students. The RTOP and Horizon instruments are also designed for rating both science and mathematics lessons, which limits the specificity with which they can ask about particular mathematical practices.

In our review of the literature and instruments, we could locate no observational measures that focused solely on the teachers' knowledge as it is used in classroom instruction; we therefore began work on our own. Building measures that would quantify actual classroom teaching turned out, as we would find, to be an intensive measures-development project of its own. Below, we orient the reader to the final shape our codes took, provide some background on their development, and supply an example of using a selection of codes to assess a videotaped lesson segment.

The MQI Instrument

Our instrument is intended to capture a range of teacher work with mathematical content, curriculum materials, and students. It consists of 83 codes grouped into five sections, and an accompanying glossary that provides overall instructions and details on each code. The five sections are:

Section I: Instructional formats and content
Section II: Knowledge of mathematical terrain of enacted lesson
Section III: Use of mathematics with students
Section IV: Mathematical features of the curriculum and the teacher's guide Section V: Use of mathematics to teach equitably

In choosing these five sections, we hoped to record not only the mathematical quality of the lesson (sections II, III, and V), but also provide information on factors that might affect mathematical quality, including particularly the mathematical content (section I) and curriculum materials with which teachers were working (section IV). We provide more detail on each section below.

Section I records the instructional format and content focus for each segment of the lesson. The instructional format code indicates the configuration of the class during this portion of the lesson, such as whether the class is working as a whole group or if students are working individually. We also note the mathematical content (e.g., number concepts, geometry, or probability) worked on in the segment—both major and minor topics. For example, in a certain

segment, students might be finding the perimeter of polygons with decimal number side lengths. In this case, we would code for both *geometry* and *operations*. Section I also captures the instructional intent of each segment: review, warm up, or going over homework; introducing the major task of the lesson; student work time; or synthesis or closure.

By themselves, the Section I codes do not reveal much about the mathematical knowledge held by a teacher. However, we found that these codes were necessary in order to make interpretations and judgments about other codes. For example, imagine a second grade lesson where the students are using base ten blocks to solve addition problems. In an introductory lesson, you might expect detailed discussion of how to use the base ten blocks to model the problem as well as explicit links between the materials and the written symbols. However, in a review lesson, you might expect the pace to be quicker, perhaps without a detailed explanation of how to set up the problems with blocks.

Section II codes the teacher's knowledge of the mathematics entailed in the lesson as revealed by its enactment. Two of the codes track on the teacher's use of language—the use of mathematical vocabulary and the general way that mathematical ideas are presented. Another set of codes in this section captures the examples and models used to represent mathematical concepts. For instance, do the examples develop the mathematics of the lesson? Are the manipulatives used appropriate models of the content? Does the teacher make explicit links between representations to highlight significant mathematical features? Does she make mathematical errors? Section II also includes three codes intended to capture different degrees of mathematical explanation: description, explanation, and justification. We expand on this in the following section. Finally, there is a global code used to record the coders' impression of the teacher's level of mathematical knowledge. Overall, this section is designed to capture the teacher's understanding of the content being taught and the mathematical resources used during the lesson.

Section III examines how the teacher uses mathematical understandings and resources <u>with students</u>. "With students" is the main distinction between Section II and Section III codes. For example, Section II captures whether the lesson segment included mathematically appropriate explanations, but does not consider who gives the explanation. Section III, on the other hand, captures whether the teacher creates opportunities for students to provide mathematical explanations, and whether students' efforts to explain are adequately scaffolded. Other codes look at how the teacher responds to students' comments, questions, ideas, or errors. For example, does the teacher correctly interpret the student's mathematical thinking, or does the response distort the mathematics or miss the point? Section III codes also include the recording of the mathematical work of the lesson and delivery of the mathematical tasks students will be working on. With these codes, we do capture some elements of teachers' pedagogical choices. However, we argue that these codes are, by and large, agnostic with regard to many current debates in mathematics education; they are intended to capture whether a teacher can work smartly both with mathematical content and students, not whether she is engaging students in a particular set of mathematical practices.

Section IV addresses the content, accuracy, and supportiveness of the lesson's curriculum materials. Our interest in curriculum materials arose from our observation that teachers were using materials, often of varying quality, in different ways. One teacher may offer more mathematical explanations or representations, for instance, because her curriculum materials support the use of such things. Or, what seems to be a teacher's mathematical error may stem from the set of curriculum materials she uses. In our final instrument we have two types of curriculum codes. The first set focuses on the curriculum's mathematical quality by assessing it

based on the codes in Section II: conventional notation, technical language, general language, and so forth. The second set asks whether the materials offer guidance for teachers on the mathematical point of the lesson: the choice/benefits of notation, language, examples and representations; details on how to work with models and representations; how students might react to the mathematics; and how to check for understanding and improve equity.

Section V captures the ways a teacher uses mathematical knowledge to teach equitably. The first code in this section consists of a two-part evaluation of any real-world problems or examples used in the segment. If a real-world context is not present, the coder marks A1 (not present). If a real-world context is present, the next step is to determine whether it is sensitive or insensitive to students' background experiences. Finally, the mathematical appropriateness of the context is evaluated: is it mathematically appropriate for the lesson's goals, or does it significantly complicate or distort the mathematics? A number of the codes in this section capture a teacher's explicitness—about the work the students are supposed to be doing, about the meaning and use of mathematical language, about ways of reasoning, and about mathematical practices. Many authors (e.g., Ladson-Billings, 1995) argue that such explicitness levels the playing field, providing instruction in these mathematical features to children who may not have been exposed to them in non-school settings. Other codes look at the opportunities students have to learn and participate in the lesson. For example, is the instructional time being spent on mathematics rather than on administrative or other concerns? Are students given opportunities to work autonomously? Does the classroom support a range of competencies and support multiple forms of mathematical contributions?

Developing the Coding Rubric: Sample and Methods

In this section, we provide information about the sample of teachers participating in our study, our data collection efforts, and the development of the coding rubric itself. By describing our sample of teachers, we hope to provide some information on the range of teachers and teaching that helped inform our code development. By explicating the development of the codes in detail, we hope to provide potential users with a history of our work and thinking about this instrument.

Sample and Recording of Practice

We recruited ten teachers to participate in our video study based on their commitment to attend professional development workshops and to participate in our study. As such, this is a convenience sample—but one that we hoped would represent a wide range of mathematical knowledge for teaching. Our teachers taught various grades from 2nd to 6th, although the 6th grade teacher was moved to 8th grade in the second year of taping. Seven of the teachers taught in districts serving families from a wide range of social, economic, and cultural backgrounds, including many non-native English speakers. For example, one elementary school within one district enrolled students speaking over 50 different languages. The three other teachers taught in the same school in a small, upper-class, primarily Caucasian district.

Teachers were taped three times in the spring of 2003 prior to a week-long mathematics-intensive professional development[2], three times in the fall of 2003, and three times again the following spring of 2004. The professional development offered five additional days of follow-up sessions in the fall of 2003. Because these teachers had all registered early for the professional

[2] Teachers attended Mathematics Professional Development Institutes in California. For more details, see Hill & Ball, 2004.

development, they might be considered unusually motivated to improve their mathematics teaching, however their scores on our measures reflect a large range (22nd to 99th percentile) in their mathematical knowledge for teaching at the beginning of the professional development.

The videotaping was done by LessonLab using high-quality professional equipment, including a separate microphone for the teacher, boom microphone for the students, and a custom-designed stand that allowed for fluid movement of the camera around the classroom. Following every lesson, teachers were interviewed about the lesson and these interviews were also videotaped. All videotapes were then transcribed by LessonLab. Following the first wave of videotaping, we realized that having copies of the curriculum the teachers used in preparing the lessons would be an important resource for analysis, so for Waves 2 and 3, curriculum materials were collected from the teachers for each of the lessons. Finally, teachers completed our pencil-and-paper measures at the beginning of the study and, for the most part, after their participation in professional development.

Developing Our Codes

The broader work of our research project seeks to identify the mathematical knowledge teachers draw upon and use during instruction, and to develop methods for accurately reporting teachers' grasp of this knowledge. Both these goals were present in developing our coding rubric. Once the first wave of videotape data became available, we began systematically listing elements in mathematical knowledge for teaching, and designing a system for "grading" particular lessons on each element. In the former task, we worked primarily from three sources: our experiences with teaching and with studying teaching and teacher education; the videotapes themselves, which we watched in small segments to help us build the codes; and the existing literature that investigates mathematical knowledge for teaching. Below, we describe three phases of our efforts to develop codes: first efforts that focused on mapping the terrain; second efforts that used literature to support existing codes that emerged in the initial design, as well suggested additional codes; and a final phase where we refined codes and developed the glossary of standard procedures and definitions. Throughout this process of video code development, as well as the actual coding, we were "blind" to teachers' scores on our pencil-and-paper measures; though we were using their tapes to help understand the nature of mathematical knowledge for teaching and to design our coding system, we did not want their performance on our measures to influence our coding scheme or appraisal of their work.

Mapping the terrain. We had three initial goals for coding: 1) to track on mathematical knowledge that appears in teaching, including agility or fluency in its use; 2) to watch for places where the teachers encounter mathematical difficulties; and 3) to develop knowledge of the mathematical issues and problems that arise in teaching. As this suggests, we wanted our codes to reflect positive uses of mathematical knowledge as well as difficulties or mistakes. We beganthis work by watching short segments of videotapes, reflecting on how teachers' mathematical knowledge appeared in the lessons, then discussing how our ideas might be translated into codes. After our first few meetings, we developed four main categories of codes corresponding to the following questions:

- What is the teacher's command of the mathematical terrain of this lesson?
- How does the teacher know and use mathematical knowledge in dealing with students?
- How does the teacher know and use mathematical knowledge in using the curriculum?
- How does the teacher know and use mathematical knowledge for teaching equitably?

These four categories remained constant throughout, and are now the foundation for Sections II-V.

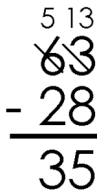
Within these broad categories, we sought to identify finer-grain elements in mathematical knowledge for teaching. Watching the videotapes suggested categories less often described in the existing literature. In some lessons, teachers evidenced considerable skill in choosing and sequencing numbers, examples, or cases to scaffold student learning. Our viewings also suggested that teachers vary in how they attend to, interpret, and handle their students' oral and written productions (e.g., students' questions in class, difficulties and confusions, innovative ideas). Examining the set of lessons also hinted that teachers vary in their ability to make connections between classroom work (following a procedure; using manipulatives) and the mathematical idea or procedure the work was meant to illustrate. Each of these initial codes was originally measured using a five-point Likert scale.

Our initial codes were revised as we attempted to use them to code more video records of teaching. The process of watching these videotapes revealed nuances and even new categories. For instance, early in our work it became apparent that teachers' treatment of mathematical language was a probable indicator of their knowledge of mathematics and a major aspect of the overall mathematical quality of a lesson. Some teachers were skillful in using mathematical language precisely, while others used less mathematically precise language, such as mis-pronouncing key terms or employing incorrect, incomplete or inaccurate language. Thus we designed a code to reflect whether teachers' "use of mathematical terminology and mathematical notation, when used, was accurate and clear." But this code proved problematic in nearly every lesson the group watched together. In some lessons, teachers primarily used non-mathematical terms to convey mathematical ideas, with different levels of skill; one teacher, for instance, taught an entire lesson on estimation without actually using this term. In many other places, teachers used mathematical terms correctly but seemed lost when trying to explain a mathematical idea or procedure to children using everyday language. And in still other places, teachers failed to differentiate between everyday and mathematical meanings for particular words (e.g., edge). This observation led us to break the language code into two:

a) Technical language (mathematical terms and concepts): Use of mathematical terms, such as "angle," "equation," "perimeter," and "capacity." Appropriate use of terms includes care in distinguishing everyday meanings different from their mathematical meanings. When the focus is on a particular term or definition, code errors in spelling, pronunciation, or grammar related to that term as present-inappropriate.

b) General language for expressing mathematical ideas (overall care and precision with language): Code general language including analogies, metaphors, and stories used to convey mathematical concepts. Appropriate use of language includes sensitive use of everyday terms when used in mathematical ways (e.g., borrow).

A similar evolution took place within another category that we anticipated being important to ascertaining teachers' mathematical knowledge in teaching: the presence of mathematical description, explanation, and justification. Defining the boundaries between these three elements of instruction, however, proved difficult. Our research team included research mathematicians, former elementary teachers, mathematics educators, and those with no formal mathematical or education background. What some saw as description, others saw as explanation—and what was seen depended, in many ways, on prior experience. A pivotal shift in coding these elements came as a result of one project member sharing an example. This project researcher explained the difference by using an example of subtraction with regrouping:



In the context of this example, mathematical description simply meant the recounting of the steps involved in subtraction with regrouping—cross out the 3, write 13, cross out the 6 and write 5. Subtract 8 from 13 to get 5…. More generally, we determined that descriptions do not address the meaning or reason for these steps. Mathematical explanations give *mathematical* meaning to ideas or procedures. In this example, the teacher (or student) might explain that the crossing-out process is really a way of re-writing the 63 as 50 and 13 ones. Re-writing in this way allows one to subtract the ones and tens column without using negative numbers.[3] Finally, mathematical justification includes deductive reasoning about why a procedure works or why something is true or valid in general. Here, the teacher might help students determine whether this algorithm can be used to subtract any two multi-digit whole numbers where trading is required. Our codes reflected this distinction:

h) Mathematical descriptions (of steps): Teacher's directing of mathematical descriptions (by self or co-produced with students) provides clear characterizations of the steps of a mathematical procedure or a process (e.g., a word problem). Does not necessarily address the meaning or reason for these steps. Code I for incomplete or unclear attempts.

i) Mathematical explanations—giving mathematical meaning to ideas or procedures: Teacher's directing of explanations (by self or co-produced with students) includes attention to the meaning of steps or ideas. Does not necessarily provide mathematical justification. Code I for incomplete or unclear attempts.

j) Mathematical justifications: Teacher's directing of explanations (by self or co-produced with students) include deductive reasoning about why a procedure works or why something is true or valid in general.

The second phase of our work involved checking our emergent coding scheme vis-à-vis the existing literature on mathematical knowledge for teaching. In a series of meetings, research staff studied selected readings with an eye towards developing new codes and revising and refining existing codes. For example, a reading of the literature around language and equity (Section V) led us to make important additions to our language codes. Mediating between students' and disciplinary language is an important element of teachers' instructional work (Pimm, 1987). Teachers' use of language, in particular mathematical terminology and meaning, is an essential resource for their students' learning of mathematics. Ladson-Billings (1995) argues that teachers must not assume that all children come equipped with the same

[3] Another example can help clarify. Teachers often teach a standard procedure for simplifying fractions—for example, to simply a fraction (such as 24/36), divide the numerator and denominator by their greatest common factor (in this case, divide both the 24 and the 36 by 12 to get 2/3). This is a mathematical description of steps; it tells you what to do, but doesn't elaborate the meaning behind the steps or provide any justification of its validity. A mathematical explanation would include a discussion of why the procedure works—that when dividing the numerator and denominator by their greatest common factor, you are, in fact, dividing by a convenient form of 1 (in this case, 12/12), and therefore not changing the value of the fraction. Still another example: In solving equations, teachers and students often make statements such as "you do the same thing to both sides." This would be marked as mathematical description. To rise to the level of explanation, this statement would need to include some sense for why this occurs – because the quantities on both sides are equal.

understandings of mathematical language; being explicit about the meaning for terms is a key component of leveling the playing field. Thus, as described by Ball, Masters-Goffney, and Bass (2005) equitable teaching includes being sensitively careful at the interface between mathematical and everyday language. This sentiment is expressed in the code V-c, “explicit talk about the meaning and use of mathematical language.”

During the process of evaluating how reliably we could apply the codes to specific lesson segments, the need for a detailed glossary became apparent. Thus, each code is accompanied by a description of the specific practices being analyzed. The glossary language was refined and revised over the course of a year for accuracy and usability.

Developing a protocol for coding and record-keeping. One of the first decisions we made during our coding trials was to determine the length of the lesson segments we would code. We quickly decided that it did not make sense to code an entire lesson at one time, primarily because the values each code took could vary across different parts of the lesson. For example, one of our codes is "development of mathematical elements of the work" and we found that some lessons had segments where the teacher was developing the mathematics of the lesson, but also had segments where the lesson got off track. It was also difficult to hold the entire lesson in mind while coding. We then tried coding only 10-minute segments, but still found it difficult to hold the entire 10 minutes in mind—and even here there were examples of codes that took multiple values even in this shorter time period. We decided that 5 minutes was a comfortable amount of time for coding, and we divided all the lessons into "clips" of approximately this length, making sure not to break the lesson at an odd place, such as the middle of a teacher or student “turn” (i.e., completed thought).[4] [5] While we coded in 5-minute segments, however, we planned to conduct analyses by aggregating codes to the lesson level, then to the teacher level. Thus our ultimate score would be a teacher-level, rather than lesson-level or segment-level score.

Even after breaking the lessons into 5-minute segments, we ran into difficulty designing a method for recording our observations. Initially, we used a Likert scale coding for each. However, the Likert scale required too many subtle distinctions to reach reliability on codes like “mathematical explanations” and “technical language.” We shifted to coding based on whether the actions indicated by the code were "present" (P) or "not present" (NP). We realized, however, that this missed two important aspects we were trying to code in sections II, III and IV: whether elements that were present were also acceptable, mathematically, or whether they contained flaws or errors; and whether elements that were absent should have occurred in order for instruction to reasonably proceed. For example, there are times when there is no need for the teacher to record the mathematical work of the lesson, such as when the teacher is handing out materials or reviewing previous work. There are other times, such as when students are analyzing data they have collected, that is essential that the work is public and available for all to see. To capture this difference, we added an additional component of "appropriate" (A) and "inappropriate" (I) to the "present" (P) and "not present" (NP) coding options for sections II, III and IV. This created four possible options for the codes in tables: Present and appropriate (P-A), present and inappropriate (P-I), not present and appropriate (NP-A), and not present and inappropriate (NP-I):

[4] The division of video for coding purposes is a difficult issue. Ideally, one would want “natural” units of instruction – a piece of the lesson that would make sense in its entirety. However, such units might be very long, and it is difficult to come to agreement on what such a unit might be for any given lesson. We elected to use 5-minute segments, and designed our codes to work at this level of analysis.

[5] We imported our videos into iMovie, in which we could divide them into roughly 5-minute clips.

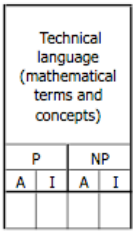


Fig. 1 below shows a coding decision tree that represents our thought process. Because we did not want to over-impose our views of "good" mathematics instruction, the glossary noted that a clip should only be coded NP-I "where the element's absence significantly hinders or obscures the mathematics, or significantly limits students' access to the mathematics in the moment" (LMT, 2005, p. 3). In this way, we hoped coders would not hold each segment to a gold standard of what *should* have happened for instruction to be ideal, but a more reasonable standard.

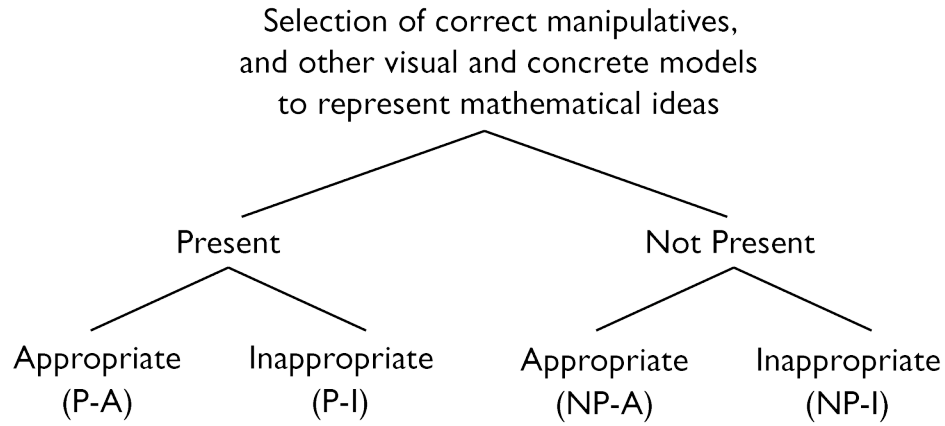


Fig. 1
Coding decision tree – example code IIe

Once we had settled the codes and method for recording the codes, we turned to trying to achieve inter-rater reliability. Our research project is composed of individuals with a variety of backgrounds and expertise; while generally beneficial, this became problematic during our reliability trials. Different persons were able to "see" different things in the lessons, and held different standards for when a mathematical element was considered present or appropriate. To address this problem, we decided to code in pairs, so that there was a broader range of expertise brought to the coding of any given clip. We developed a protocol wherein random pairs[6] were assigned to code each lesson. In order to get a sense of the mathematics and its development, coders watched the entire lesson before beginning coding. Then, each coder coded the lesson clip by clip. When necessary, coders could reference transcripts, interviews, and for many of the lessons, curriculum materials. Coders then met and reconciled their codes before making their record permanent. Much of the time coders agreed and only needed to discuss a small number of codes for which they had initially differed. We conducted several rounds of trial coding by pairs as we finalized the codes and glossary. At the end of this

[6] Alternatively, we could have assigned coders based on differences in background and expertise.

process, an official interrater reliability check (using the codes produced by pairs as the unit of analysis) computed the final rates of agreement by code and by table. Agreement ranged between 65 to 100% for individual codes and averaged 89 to 90% for entire tables (II, III, and V). Finally, coders wrote a short synopsis of the lesson and its mathematical strengths and weaknesses for our records.

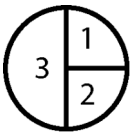
Coding of the 90 lessons took place in pairs over a period of roughly six months. Curriculum materials were independently coded by an individual whose scholarly research focuses on teacher learning from curriculum materials. Intercoder reliability was not computed.

Examples of Selected Codes in Use

To further elaborate our codes and illustrate how they can be used to capture teacher knowledge in use, we will share episodes from two different lessons. One lesson is from a second grade class and the other from fourth grade; both lessons took place in the spring, at the end of the school year. In each lesson, students were using spinners to investigate concepts in probability. We first give a brief description of each of lesson, and then explain how our language and explanation codes capture aspects of each lesson.

In the second grade lesson, students worked with spinners that were divided into equal-sized segments (e.g., circles marked off into thirds, fourths, or sevenths). After making sure that all students in a group had a different type of spinner, the teacher (Teacher S) directed students to each color in one of the fractional sections of their spinner. Teacher S explained the experiment, demonstrating how to record a tally for "yes" if the spinner landed on the colored segment, and "no" if the spinner landed on any of the white segments. The teacher then told the class to begin working individually, spinning and recording their own data. As the students worked, the teacher circulated among the students, answering questions and helping students. Once everyone had recorded the outcomes of their 10 spins, the teacher had students take turns holding up their spinners and sharing their results (e.g., 8 noes and 2 yeses; or 1 yes and 8 noes). Teacher S then asked the students to color in two additional segments on their spinner (i.e., three segments would now be colored) and repeat the experiment. The lesson concluded with the teacher asking students to raise their hands if they got certain results in their second experiment, such as if they had three or more yeses, and then asked the class what they had learned from this work.

At the beginning of the fourth grade lesson, the teacher (Teacher F) distributed spinners to the class and reviewed their use. The spinners were divided into three sections—labeled 1, 2, and 3—with Sections 1 and 2 each taking one-fourth of the spinner's area, and Section 3 the remaining half:



In the first experiment conducted during this lesson, the students spun their spinners and recorded the outcomes on their graph. They continued spinning and recording outcomes until one of the numbers reached the top of their graph. Before beginning the experiment, Teacher F introduced the term "outcomes" and asked students to make written predictions in response to the following questions:

1. When you spin the spinner, is any number more likely to come up than any other number? Why do you think so?
2. What do you think the entire graph will look like when one number reaches the top of the paper?

After students posted their results, the class discussed their predictions and outcomes, and students offered explanations for why they thought 3 would be the most likely outcome. Near the end of the lesson, the class began a second experiment involving the sum of two spins that would be concluded in the next day's lesson.

Mathematical Language Codes

Three of our codes capture a teacher's use of mathematical language during a lesson: *technical language* and *general language for expressing mathematical ideas* in Section II, and *explicit talk about the meaning and use of mathematical language* in Section V. *Technical language* reflects a teacher's use of mathematical terms, and *general language* focuses on the teacher's overall care and precision in discussing mathematics. *Explicitness* goes beyond just using terms and language, requiring a sort of "meta-discussion" about mathematical language—for example, defining terms, discussing how to use notation or the meaning of symbols, or pointing out specific labels or names. Examples of some these different types of language use can be seen in the following segment from the fourth grade probability lesson:

Teacher F: In spinning the spinner, what are the possible numbers the spinner could land on when spinning it? What are the possible numbers? Razi?
Razi: 1, 2 and 3?
Teacher F: Okay. 1, 2, and 3. Any other possible numbers?
Razi: No.
Teacher F: Yeah, okay, that's it. You know what those numbers are called? There's a name for those numbers. They're called "outcomes." Okay? The possible results of an experiment are called "outcomes." So how many outcomes does your spinner have?
Students: Three.
Teacher F: Okay. Your spinner has three outcomes. Those are all of the possibilities that exist for spinning the spinner and predicting which number it's gonna land on.

In this exchange, the teacher correctly used "outcomes," a mathematical term, to name the possible results of an experiment. We would, therefore, code "present and appropriate" (P-A) for *technical language*. Furthermore, the teacher did not just use the term "outcome" correctly; she explicitly discussed its definition and asked students to use it to describe a particular spinner. In addition, though not visible in the transcript above, after introducing the term, the teacher displayed a poster on the chalkboard that listed the term "outcome" with its definition. Thus, for this segment, we would also code "present" for *explicit talk about the meaning and use of mathematical language*.

This explicitness about language can be contrasted with an example from later in the same lesson where the fourth grade teacher also correctly used the term "outcome" in her introduction of the second experiment, but did not explicitly discuss its meaning or use:

Teacher F: Who can share with me? What are the possible sums that we could get? What are the possible outcomes or sums that we could get by adding two spins together? Zack?

We would code this use of "outcome" as "present and appropriate" for *technical language*, but as "not present" for *explicitness*.

Teachers, however, do not always use mathematical terms precisely. Consider the following example from the second grade lesson. The class had already completed the first spinner experiment, in which one section of their spinners was colored. Teacher S then directed the students to color in two additional sections, "Okay. Let's do this. Everybody color in two more squares. Color in two more squares." We would code this example as "present and inappropriate" (P-I) for *technical language.* The term "square" is a mathematical term that is present in the episode. However, the sections of the spinners are not squares; therefore the use of the term is inappropriate. Although one instance of an inappropriately used term is enough to warrant a (P-I), this teacher inaccurately refers to the sectors as "squares" a total of five times throughout the segment.

Present and appropriate (P-A) for *general language* is more difficult to demonstrate with a small segment of transcript. For a clip to be coded as (P-A), the teacher must discuss mathematical ideas carefully and precisely throughout the entire five minutes. However, like *technical language,* one instance of sloppy or imprecise general language results in a (P-I). In a different lesson, for instance, a third grade teacher had students weigh different objects from their desk (pencils, erasers) on a pan balance, using blocks as counterweights. Throughout, her description of the point of the lesson and actual mathematical task was garbled, as in this excerpt from the end of the lesson:

Okay. Stop. All right. Huh, the lesson today was just about practicing weighing something, okay? It's not like when you get on the scale at home and it goes right to a number. This is actually taking different weights and trying to get it to equal evenly.

In fact, she used the terms "equals out" and "equals evenly" rather than "balance" throughout the lesson. The "different weights" in the last sentence has an unclear referent – is she talking about the blocks used as counterweights, or the different weights of the objects from students' desk? Although this teacher did not use any technical mathematical language incorrectly, this is a clear instance of poor use of general language to describe mathematical ideas or procedures.

Mathematical Explanation Codes

Section II contains three codes related to explanation: *mathematical descriptions*, *mathematical explanations*, and *mathematical justification*. In addition, Section III captures whether students are involved in the giving of mathematical explanation with two codes: *elicits student description* and *elicits student explanation*.

The following excerpt from the fourth grade probability lesson contains examples of mathematical explanation. The class had completed the first spinner experiment, recording outcomes of single spins of the 1-2-3 spinner, and found that 3 was their most frequent outcome. The teacher asked whether their initial predictions matched the class results:

Teacher F: Okay. Well, comments? How many of you predicted number 3 would come up the most often? [Most students raised their hand.] Wow. Who can tell me why? Why you made that prediction? Zack?

Zack: I made that prediction because on the spinner, the 3 has half of the spinner, and 1 and 2 only have a quarter of it.

Teacher F: Okay. So 3 has half the spinner—3 had half the spinner, and then, 1 had one-fourth, and 2 had one-fourth of the spinner. Okay. Anybody else—comments about why you predicted number 3 would come up the most often? Ghadah?

Ghadah: On the spinner, 3 had a bigger shape.

Teacher F: What do you mean by a bigger shape? Can you explain that?
Ghadah: It was like wider than 1 and 2.
Teacher F: All right. Missy?

In this episode, the teacher elicited explanations from the students for why they predicted that 3 would be the most frequent outcome of the experiment. We would code this as "present and appropriate" (P-A) for both *mathematical explanation* and *elicits student explanation*. If, instead, only the teacher had provided an explanation for why 3 is the most likely outcome, then it would be coded as "present and appropriate" (P-A) for *mathematical explanation*, but "not present and appropriate" (NP-A) for *elicits student explanation*.

There are also cases when a teacher elicits student explanation, but the student's response does not provide sufficient explanation of the meaning of a mathematical idea. This occurred during the second grade lesson. To demonstrate how students were to report the data from their spins, the teacher asked a boy near the front of the class to spin his spinner three times. Each spin landed on a white segment of his spinner, and the teacher recorded three tallies next to "no." She then asked the class what these results might imply about the student's spinner:

Teacher S: What's your thought so far? Do you think he has many divisions in his or do you think he has few?
Charles: Many.
Teacher S: Many because of what? Why, Charles?
Charles: There's three noes and zero yeses.
Teacher S: Good boy. Okay. Do you understand what to do? When he lands on the colored, he's going to put a tally on the yes. You land on the color part. If you don't land on a color part, you're gonna put no. Ready? Go.

In asking Charles why he thought there were many divisions, the teacher elicited a student explanation. However, Charles simply restated the results ("There's three noes and zero yeses."), which is not a mathematical explanation for why those results would imply that the spinner was divided into many segments. The teacher did not scaffold Charles' explanation with further questioning, nor did she provide an explanation of her own. It was instead a non-explanation that was accepted as an explanation. We would code this segment as "not present and appropriate" (NP-A) for *mathematical explanation* because there was no explanation (NP), and yet we could not make a strong case that an explanation was *required* at that moment in instruction. However, this segment was coded "present and inappropriate" (P-I) for *elicits student explanation* because a student explanation was elicited, but not supported.

The conclusion of the second grade lesson provides another example of mathematical explanation, and illustrates one important issue with our coding effort. At the end of this lesson, the teacher asked the students to reflect on the day's activities and to explain how the spinner experiments related to an earlier probability experiment that involved drawing colored cubes out of socks:

Teacher S: So what did we learn from this? What did we learn from the spinners in relating it also to the socks? What did we- how can we– how could we correlate- how could we figure this- how could we put this together? What can we–?
Lily: The more of the same thing, it's more likely that you'll pick it.
Teacher S: Do you agree with that?
Students: Yeah.
Teacher S: Can you say it any other way? Anybody want to say it differently?
Teacher S: Nicely done, Lily. Good job.

While this is a nice start to an explanation of the mathematical goal of the lesson or to a discussion about probability, the teacher did nothing to probe or help elaborate the student's

observation. She asked the class if anyone else would like to state the idea another way, but when no one raised their hand, she ended the lesson with no further explanation. Although there were no other instances of mathematical explanation during the five-minute segment, we would code this exchange as "present and appropriate" (P-A) for both *mathematical explanation* and *elicits student explanation*, because a single instance of explanation is enough to warrant a (P-A) for both of these codes.

The crux of the problem is, as the reader might have guessed, this: when a five-minute segment of video is coded as (P-A) for *mathematical explanation*, it could represent a single instance of underdeveloped explanation, as in the second grade example above, or it could be the result of a five-minute segment containing multiple, highly detailed explanations—it is impossible to tell from the code alone. Unlike the codes for inappropriate mathematical language, computational errors, and other clearly incorrect moments in instruction, where a single instance led to a code of "inappropriate," coding positive examples of mathematical knowledge used in instruction led to lengthy group discussions about where to place the "bar." Does an explanation have to be fully developed before it counts? We imagine that there are readers who would argue that neither the first nor the third example rose to the level of a true mathematical explanation; there are some who would count the first but not the third as an explanation; and there may be some that count both. Coming to agreement on where to place the bar was one important component of our efforts to develop these codes.

As the above examples suggest, we erred on the side of placing the bar low for positive uses of mathematical knowledge for teaching, counting as "explanations" or "justification" lesson segments that barely rose to the glossary definitions for these terms. We expected, however, that the process of aggregating across lesson segments and lessons to arrive at a teacher's overall "score" would ultimately accurately reflect a teacher's knowledge in this area, and we were correct. The second grade teacher featured mathematical explanations in only 8% of her clips; the fourth grade teacher featured mathematical explanations in 37% of her clips. This gap would likely persist no matter where we placed the "bar" as long as that bar was consistently applied.

We believe many of our codes will encounter this "bar" problem, and discuss some implications in the conclusion.

Analysis

Our initial analysis plans focus on the correlation of teachers' pencil and paper measure scores with their videotape "scores." This allows us to examine the predictive validity of the multiple choice measures: Do they in fact predict the mathematical quality of teachers' classroom work? We are now pursuing three types of analysis.

The first and most complex avenue consists of data reduction along major themes that emerged from watching the tapes. By compressing similar codes into what we now call "meta-codes" and aggregating scores to the lesson and teacher level, we can examine how several major dimensions of mathematics instruction relate to performance on our measures. These dimensions include whether mathematics is taking place during mathematics instruction; the presence of mathematical inaccuracies and errors; whether and how well teachers respond to student thinking and errors; and the presence of "rich" mathematics. This analysis is now under way, and we expect results within a year.

The second avenue compares teachers' overall lesson and pencil-and-paper scores. Overall lesson scores were assigned by the pair of coders after watching the entire lesson; each lesson was assigned a "low," "medium," or "high" score, based on our observation of the mathematical knowledge for teaching displayed in that lesson's instruction. Averaging these scores for each teacher and comparing with her pencil and paper measure score yields a correlation of .79.

Results from the third analysis are also available. Prior to revealing teachers' measure scores, the research group rank-ordered teachers based on our overall perceptions of the mathematical knowledge they exhibited in teaching. This rank-ordering turned out to be strongly predictive of their pencil-and-paper measure scores: two teachers were placed exactly correctly, four were off by only one, and no teacher among the four remaining were off by more than three places. The Spearman rank-order correlation between teachers' projected and actual rank was .79 ($p < .01$), and the correlation between their projected rank and actual score on the pencil-and-paper measures was -.77 ($p < .01$). Although a small dataset and analysis, some patterns pose intriguing hypotheses for future research. For instance, teachers in the upper half of the ranking were nearly exactly correctly predicted, while teachers in the lower half of the ranking distribution were more likely to have had their position mis-predicted by the group. This suggests either that our measures are less reliable at the lower end of the distribution of mathematical knowledge for teaching, or that the relationship between mathematical knowledge and mathematics teaching is non-linear—that for teachers below a certain criterion (roughly 2/3 correct on our measure) the lack of mathematics knowledge causes more error-filled, less meaningful instruction, period.

Whatever the case, however, the correlations between teachers' pencil-and-paper score and videotaped ranking/overall score is quite high. This strengthens our view that the pencil-and-paper assessments capture important knowledge for teaching mathematics, and helps illuminate the ways in which teachers' mathematical knowledge base can be put to use in classrooms.

Conclusion

In this conclusion, we consider some issues we anticipate will arise as other groups begin to use the measures. Next we consider the dissemination of these codes. One major underlying issue is that the coding rubric itself, and the glossary that goes with it, rely on words to convey meaning. Yet words cannot adequately convey what we mean by a code because of variation in the meaning of terms; what one person views as "explicit talk" about mathematical language or reasoning will be viewed differently by another. How explicit is explicit? Although we constructed a glossary to help alleviate this problem, even casual users will recognize that it is inadequate. Rather, we believe that the extent of intercoder reliability we reached, as a project, was driven in large part by the two years of conversations we had around the development of the instrument itself. We may not all agree on the finer points of "classroom work is connected to mathematical idea or procedure," but we had all watched the same videotapes—and participated in the same discussions—that provided concrete instantiations of this *not* happening. By benchmarking lessons in this way, we developed an understanding of how to code videotape that went beyond what we could write down in the glossary or the codes themselves. This understanding cannot easily be transmitted to other research projects without developing training materials.

A related point is that for many codes, where to set the "bar" between appropriate or inappropriate (or present/not present) was a matter of group-developed norms and conventions. We all clearly agreed, for instance, on a few segments where classroom work was *not*

connected to a mathematical idea or procedure. There were many segments where classroom work was connected. But in between these clearer examples, there was a gray zone—segments where members of the group had legitimate yet different opinions. Resolving these gray zone cases—as we did over lengthy discussions on Monday mornings—allowed us to "set the bar" for many codes. But again, we cannot replicate these discussions in our codes or glossary. As a result, we expect that researchers using these codes will need to spend considerable time going through the same trial codings and discussions before reaching agreement.

Third, we strongly suspect that coders must possess high levels of mathematical knowledge, and knowledge of mathematics for teaching, to code accurately. Even within our group, we found that different individuals "saw" different strengths and weaknesses of teaching based on their background. Our mathematician, Hyman Bass, often pointed out subtle mathematical errors and mis-steps that the rest of the group missed. And our expert mathematics educators often pointed out instructional problems that others (including Hyman) missed. In one example, for instance, one teacher was elegantly explaining the long division algorithm by linking steps in the standard procedure to manipulatives—but switched between the partitive and measurement interpretation of division as she did so. Only one mathematics educator noticed this problem of interpretation, but once she explained her thinking, the group agreed this was a significant problem for the lesson segment. Our intuition is that "training" observers in the knowledge it takes to code these tapes is likely to be of little use; instead, observers with strong knowledge in both these arenas must be found.

We have also seen that the variation in mathematics instruction is immense. Although we feel confident that we have captured some of the major ways mathematical knowledge is used in teaching, there are doubtless others that our codes do not cover. What to do when this arises is one dilemma research projects adopting this measure face.

Finally, this instrument must be adapted for use in real-time classroom situations, where coders do not have the luxury of replaying events, consulting transcripts, or studying curriculum materials. We imagine that adaptation is possible, but would require a reduction in the number of codes in play during any one observation, and a staggered schedule (2 minutes on, 2 minutes off) for observing.

References

Appleman, D., & Thompson, M. J. (2002). "Fighting the toxic status quo": Alfie Kohn on standardized tests and teacher education. *English Education 34*(2): 95-103.

Ball, D. L., & Bass, H. (2003). Making mathematics reasonable in school. In G. Martin (Ed.), *Research compendium for the Principles and Standards for School Mathematics* (pp. 27-44). Reston, VA: National Council of Teachers of Mathematics.

Ball, D.L., Hill, H.C. & Bass, H. (2005) Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator*, Fall 2005, 14-22.

Ball, D. L. Masters-Goffney, I., and Bass, H. (2005). The role of mathematics instruction in building a socially just and diverse democracy. *The Mathematics Educator*, 15(1), 2-6.

Berliner, D. (2005) The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*(3), 205-13.

Borko, H., Eisenhart, M., Brown, C., Underhill, R., Jones, D., & Agard, P. C. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education, 23*(3), 194-195.

Hill, H. C. & Ball, D. L. (2004) Learning mathematics for teaching: Results from California's Mathematics Professional Development Institutes. *Journal of Research in Mathematics Education 35*, 330-351.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Horizon Research (2000). Inside the classroom observation and analytic protocol. Chapel Hill, NC: Horizon Research, Inc.

Ladson-Billings, G. (1995). Making mathematics meaningful in multicultural contexts. In W. G. Secada, Fennema, Elizabeth, Byrd Adajian, Lisa (Ed.), *New directions for equity in mathematics education* (pp. 126-145): Cambridge University Press.

Leinhardt, G., & Smith, D. A. (1985). Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology, 77*(3), 247-271.

Learning Mathematics for Teaching (2005). *Video coding glossary*. Ann Arbor: University of Michigan.

Learning Mathematics for Teaching (2006). *Measures of Mathematical Knowledge for Teaching*. Ann Arbor: University of Michigan.

Pimm, D. (1987). *Speaking mathematically: Communication in mathematics classrooms*. London: Routledge.

Sawada, D. & Piburn, M. (2000). *Reformed teaching observation protocol (RTOP).* Arizona State University: Arizona Collaborative for Excellence in the Preparation of Teachers.

Sherin, M. G. (1996). *The nature and dynamics of teachers' content knowledge*. Unpublished doctoral dissertation, University of California Berkeley.

Thompson, P. W., & Thompson, A. G. (1994). Talking about rates conceptually, part 1: A teacher's struggle. *Journal for Research in Mathematics Education, 25*(3), 279-303.