Characterizing Mathematics Classroom Practice: Impact of Observation and Coding Choices

Marsha Ing, University of California, Riverside, and Noreen M. Webb, University of California, Los Angeles

Large-scale observational measures of classroom practice increasingly focus on opportunities for student participation as an indicator of instructional quality. Each observational measure necessitates making design and coding choices on how to best measure student participation. This study investigated variations of coding approaches that may be feasible in large-scale studies, and the ramifications of these variations for drawing inferences about instructional quality. Using data from classroom videos, we found that decisions about whether to keep track of individual students in the coding, observe multiple contexts in the classroom (e.g., whole-class and small-group discussions), and capture nuances of student participation changed the resulting characterizations of classroom practice. Most importantly, simplifying the coding approach did not fully capture and even misrepresented the level and nature of student participation in many classrooms.

Keywords: classroom observations, instructional quality, mathematics

There is great interest in collecting large-scale observational measures of instructional practice. The Gates Foundation, for example, currently sponsors the collection and analysis of classroom-video observations from over 3,000 teachers four times per year (Bill and Melinda Gates Foundation, 2009). In addition, all of the winning applicants in the 2010 Phase 2 Race to the Top Fund competition (Florida, Georgia, Hawaii, Maryland, Massachusetts, New York, North Carolina, Ohio, Rhode Island, and the District of Columbia) that competed for over $4 billion in federal funding included some sort of classroom observation component in their grant application (U.S. Department of Education, 2010).

The intense focus on collecting observational data is due in part to the potential of this type of data to provide comprehensive information about instructional quality. This information can be used for various purposes, such as improving teachers’ effectiveness (e.g., through feedback provided to practicing teachers, or as information incorporated into education for beginning teachers) and distinguishing among teachers for accountability purposes. This paper focuses primarily on the former purpose: improving teacher effectiveness.

This study focuses on one aspect of instructional quality—the nature of student engagement in mathematics classroom discussions (both whole-class and small-group) — that has been shown in small-scale studies to predict student achievement. The specific issue addressed here is how classroom observations of student participation can be carried out on a large scale to provide useful feedback for improving teachers’ effectiveness. In particular, we examine the impact of various decisions about the design of observational measures (how many students to observe, how many classroom contexts to observe, whether to keep track of individual students in the coding, and whether to conduct intensive coding of video versus transcripts or observations that do not require such intensive coding) on the information that is produced about the student participation that takes place in teachers’ classrooms.

The increasing attention being paid to student engagement as an indicator of instructional quality reflects numerous and long-standing calls from educators, policymakers, and researchers for teachers to move away from purely teacher-centered instruction and move toward involving students as more active participants in classrooms (Cobb, Stephan, McClain, & Gravemeijer, 2001; Hiebert & Grouws, 2007). These calls are based on research showing the benefits for student learning when students interact with the details of the content in collaborative group settings (e.g., Brown & Palincsar, 1989; Fuchs et al., 1997; Howe et al., 2007; King, 1992; Nattiv, 1994; Saxe, Gearhart, Note, & Paduano, 1993; Slavin, 1987; Veenman, Dennesen, van den Akker, & van der Rijt, 2005; Yackel, Cobb, Wood, Wheatley & Merkel, 1990). Not all student participation or engagement leads to productive student outcomes. In detailing the level of student participation and engagement, specific feedback to teachers on ways to improve the effectiveness of their classroom practice can be generated.

There are several ways of describing student participation and engagement that relate to student outcomes. For example, explaining one’s thinking can help students internalize principles, become aware of misunderstandings and lack of...
understanding (Chi, 2000; Cooper, 1999), reorganize and clarify material in their own minds, fill in gaps in understanding, acquire new knowledge, and develop new perspectives and understanding (King, 1992; Peterson, Janiecki, & Swing, 1981; Rogoff, 1991; Saxe et al., 1993). Recent research examines this relationship in the whole-class context as well (Franke et al., 2009).

Previous research shows a number of observational approaches for obtaining information about student engagement that may be helpful for improving teaching quality (e.g., Kilday & Kinzie, 2009). An indirect approach is to draw inferences about student participation from the kinds of questions that teachers pose, such as requests for students to recall facts or describe problem-solving strategies (e.g., Atkins, 1999; Friedman, 1976; Gall, 1970, 1984; Hatano & Inagaki, 1991; Hedrick, Vernon-Peagans, & Ginsberg, 2010; Hiebert & Wearne, 1993; Hufferd-Ackles, Fuson, & Sherin, 2004; Nathan & Kim, 2009; Rittenhouse, 1998). For example, an observational protocol, Mathematics Quality of Instruction (Hill et al, 2008; Learning Mathematics for Teaching, 2006, 2010), uses this indirect approach and analyzes videos of teacher–student interactions to code whether the teacher elicits student mathematical explanations by asking students to provide an explanation or justification for a procedure.

A more direct observational approach is to code participation of students themselves. One version of a direct approach is to code student participation during teacher–student interactions. For example, Kawanaka and Stigler’s (1999) video analysis of discourse in 231 classrooms in Germany, Japan, and the United States distinguished between student utterances in terms of the amount of time and the quality of the utterances. More specifically, they distinguish between utterances made in response to teacher elicitation or direction, those intended to elicit an immediate communicative response from the teacher or from other students, and those intended to acknowledge or evaluate another student’s response. This video study of instructional practices was prompted by earlier results showing achievement differences between countries.

Although Kawanaka and Stigler’s (1999) observational approach codes student participation only in the context of conversations between teacher and student, other observation protocols include student participation that occurs apart from teacher–student interaction in addition to student participation with the teacher. A recent study observed and coded student participation both during whole-class discussions and student-directed conversations (Franke et al., 2009). That study showed, first, that a particular dimension of student participation—providing correct and complete explanations of their mathematical thinking—was positively correlated with students’ mathematics achievement. In contrast, a more global dimension of student participation—providing explanations (without regard to accuracy and completeness)—was not as strongly predictive of students’ achievement. Second, student participation differed substantially across classrooms, with some classrooms exhibiting high rates of high-quality student explaining and others exhibiting much lower levels.

The coding of student participation in the study just described was based on observations of multiple students in multiple classroom contexts and intensive analysis of classroom videotapes and transcripts, an approach that is probably not feasible for use in large-scale studies. The study presented here investigates variations of this observation/coding approach that may be feasible in large-scale studies, and the ramifications of these variations for drawing inferences about teaching quality. Specifically, we examine the following dimensions along which the observation approach may vary: intensive videotape/transcript coding versus once-through viewing of videotapes (which may simulate live observations); whether or not the coding keeps track of individual students; observation of the same classroom on one or multiple occasions; observing whole-class and small-group contexts versus only the whole-class context; and whether or not explanations are coded with respect to quality (accuracy and completeness).

### Method

#### Sample

This study reports information on student participation from six second and third grade classrooms in a large urban school district in Southern California. The schools are large, serve predominantly Latino (over 70%) and African American (approximately 30%) students, have a high percentage of students receiving free or reduced lunch (approximately 90%), have a substantial proportion of English Language Learners (approximately 50%), and have low standardized achievement scores. Class size and teacher experience were similar across schools and classrooms. The 44 students included in this study were randomly selected from these six classrooms (80% Hispanic and 20% African American).

#### Classroom Instruction

In the six classrooms analyzed here, teachers taught problems of their choice related to the topics of equality and relational thinking. Teachers posed such problems as (a) $50 + 50 = 25 + \Box + 50$, and (b) $11 + 2 = 5 + 8$ (true or false?). Further details appear in Webb et al., 2008). Consistent with their accustomed practice, most teachers incorporated group-work time into the class wherein students worked together to solve and discuss problems assigned by the teacher. Typically, teachers introduced a problem, asked small groups (often pairs) to work together to solve the problem and share their thinking, then brought the whole class together for selected students to share their answers and strategies with the whole class, usually at the board. Teachers varied in how much time and how many problems they devoted to small-group work. One teacher did not use small-group work at all, and engaged with students during whole-class discussions exclusively. Table 1 gives the number of problems each teacher discussed with the whole class and the number of opportunities each teacher provided for small groups to discuss problems across the two days of observation.

#### Recording Procedures

We videotaped each of the six classrooms on two occasions using multiple video cameras and audio setups to capture

---

**Table 1. Number of Problems**

<table>
<thead>
<tr>
<th>Classroom</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of problems discussed in the whole-class context</td>
<td>13</td>
<td>8</td>
<td>8</td>
<td>30</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Number of problems assigned to small groups to discuss</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>13</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 2. Examples of Student Explanations

<table>
<thead>
<tr>
<th>Problem</th>
<th>Student Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct and Complete</td>
<td>10 + 10 − 10 = 5 + □</td>
</tr>
<tr>
<td></td>
<td>Five! It's 'cause look. We could do this, oh no. Hold up. 'Cause ten plus ten equals twenty, huh? And then it says minus ten equals five plus blank. So it gotta be equal ten, so five plus five equals ten. And that's how I got it. Get it?</td>
</tr>
<tr>
<td></td>
<td>20 + 10 = 10 + □</td>
</tr>
<tr>
<td></td>
<td>Twenty plus 10 is 30, so...the equal sign means that you have to be the same, it has to be the same, so if there's a 10 here, and a 20 has to be there. Twenty plus 10 is 30, 10 plus 20 is 30.</td>
</tr>
<tr>
<td>Ambiguous or Incomplete</td>
<td>8 + 2 = 7 + 3 (True or False?)</td>
</tr>
<tr>
<td></td>
<td>[True] because there's a two and a three and a seven and an eight. They're like an order. [Although the answer is correct, this explanation does not make it clear (a) whether the student is considering the difference in quantity between 2 and 3 and between 7 and 8, (b) what the student means by “order,” and (c) how the student is using “order” to justify that the number sentence is true]</td>
</tr>
<tr>
<td></td>
<td>100 + □ = 100 + 50</td>
</tr>
<tr>
<td></td>
<td>The 50 will go right there because it has to be the same number. [Although the answer is correct, this explanation does not make it clear whether the student is using “equality” to justify that the 50 is the missing number or whether the student thinks there should be an equal number of numbers on both sides of the equal sign]</td>
</tr>
<tr>
<td>Incorrect</td>
<td>4 + 9 = 5 × 3 − 2</td>
</tr>
<tr>
<td></td>
<td>(True or False?)</td>
</tr>
<tr>
<td></td>
<td>10 + 10 − 10 = 5 + □</td>
</tr>
<tr>
<td></td>
<td>I thought it was false because four plus nine is thirteen, and five times three is fifteen. Those two do not match.</td>
</tr>
<tr>
<td></td>
<td>Because there's a 10 + 10. And if you had a plus sign and that would be 30, but there is a take away sign. So 10 + 10, take away 10 equals 20 plus 10 plus 10 equals 20. And then, um, 5 plus, I knew it was 15 because I went, um, I counted—5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20—and then it was 15, then I put 15.</td>
</tr>
</tbody>
</table>

student and teacher talk throughout the class session. The procedure for video and audio recordings was approved by the Institutional Review Board. These procedures included parental and student consent for student images to be used for research. The analyses presented here focus on eight randomly selected students within each class who were captured on videotape. Because of equipment malfunction, the talk of four of the videotaped students was inaudible, yielding a total sample of 44 students available for video analysis.

The eight students in each class were captured on two video cameras. One of the video cameras had a wide-angle lens, which allowed us to capture the entire classroom on video. Both the video cameras had two audio feeds (connected to flat microphones taped to the students’ tables). These two feeds per camera allowed us to capture video and audio data for two pairs (four students) per camera (eight students total).

**Coding**

*Rater training.* Before coding, raters were given instruction about how to code student explaining. We defined explanations as verbal descriptions that conveyed information about the strategy a student used to solve a problem, and included more than just the answer to part or all of the problem. Raters reviewed examples illustrating the difference between a student providing an answer and a student providing an explanation, such as the following student responses for the problem, 14 ÷ 2 = (3 + □) + 1: “The answer is two” (answer) versus “Because seven times two is fourteen so that equals seven this side has to equal seven. So, three times a number plus one. So, three times two equals six plus one equals seven” (explanation). Table 2 provides additional examples of explanations that were provided to raters to help them differentiate between explanations that were correct and complete and explanations that were incorrect, incomplete, or ambiguous.

**The coding protocols.** We used two coding protocols. One kept track of individual students’ participation (student-based protocol) and the other coded student participation on problems without regard to which students participated (problem-based protocol). We used the two protocols because they potentially yield different information about student participation in classrooms. The student-based protocol makes it possible to gauge the typical level of participation of an individual student. However, it does not provide information about whether student explanations are concentrated in the same problem or spread out across problems.

The problem-based protocol reveals whether student explanations occur in few or many problems. It cannot be used to infer the participation of an individual student, however. For example, a large percent of problems with student explaining could occur: (1) when only one student in the class offers an explanation but offers an explanation on every problem; (2) when a different student contributes an explanation on different problems; or (3) when multiple students in the class contribute explanations on multiple problems. In each of these scenarios, students in the class have an opportunity to hear explanations on many problems, but students’ individual contributions differ across scenarios. For each protocol, to account for the different number of mathematics problems across classrooms (see Table 1), we divided the total frequency by the number of problems in a classroom yielding the number of explanations per problem.

**Coding procedures.** Two raters watched the videotapes and scored them using the two coding protocols. For both the protocols, to simulate a live observation, raters watched each video once. For the student-based protocol, raters were provided a grid for each classroom with an alphabetical list of the eight students’ first names organized by video/audio source, a brief statement of the problems covered each day, and a seating map. Because there were two video sources
per classroom, each with two audio sources, there were four video/audio sources for raters to code. All four sources captured the whole-class discussion, but each one captured the discussion of a different small group. Consequently, raters coded one video/audio source in its entirety (whole-class interaction plus the interaction in one small group) and coded only the small-group conversations on the other three audio/video sources.

Raters were instructed to watch the videos (or the relevant video portions) without pausing or stopping, and to mark on the grid whenever a student provided an explanation for a particular problem and whether it was correct and complete. Raters coded which student provided an explanation during whole-class discussion based on information provided by the teacher. Teachers almost always called students by their first names during the whole-class discussion, or referred to the student who shared the answer or strategy when engaging with other students during whole-class discussion. The raters used all of this information during whole-class discussion to determine which student provided the explanation. Raters identified which student provided an explanation during small-group work based on information provided by the video and by the seating chart that an observer had sketched at the beginning of class. Raters could see from the video which student was talking and matched this information with the seating chart.

For the problem-based protocol, the coding grid for each classroom listed the problems covered on each day but did not list students’ names. Raters marked whether an explanation was given on a problem and whether it was correct and complete. It is possible that a single student provided multiple explanations per problem. In the case of multiple correct, complete explanations given for a problem, raters marked only one correct, complete explanation.

Coding variables. The metric of the student-based variables was the percent of problems for which a student gave an explanation. We formed four variables using this metric: the percent of whole-class problems on which the student gave a correct/complete explanation; the percent of whole-class problems on which the student gave any explanation (regardless of accuracy or completeness); the percentage of total problem opportunities across both the whole-class and small-group contexts in which the student gave a correct/complete explanation; and the percentage of total problem opportunities across both the whole-class and small-group contexts in which the student gave any explanation. The variables pertaining to the whole-class problems only include the problems where the teacher provided opportunities for the whole class to discuss the problem. The variables pertaining to the total problem opportunities include problems where students had opportunities to discuss as a whole class and within small groups (Table 1).

The metric of the problem-based variables was the percent of problems for which an explanation was offered. The four variables based on this metric were: the percent of whole-class problems on which a correct/complete explanation was provided; the percent of whole-class problems on which any explanation was provided (regardless of accuracy and completeness); the percentage of total problem opportunities across both the whole-class and small-group contexts in which a correct/complete explanation was provided; and the percent of total problem opportunities across both the whole-class and small-group contexts in which any explanation was provided. For the variables that included both whole-class and small-group discussion, the number of problem opportunities was the sum of the number of problems discussed in the whole class and the number of problems assigned to small groups.1

Coder consistency. We conducted generalizability analyses to examine interrater consistency (Brennan, 2001; Shavelson & Webb, 1991). For the student-based protocol, the design of the generalizability analyses was students crossed with raters. Because interest lay mainly in identifying the absolute level of student explaining (not only comparing students against each other), we present dependability coefficients (that provide information about differences between raters on the absolute levels of student explaining) in addition to generalizability coefficients (that pertain to rater differences in the relative standing of students). For the problem-based protocol, the design of the generalizability analyses was classrooms crossed with raters. Again, we present both dependability coefficients (pertaining to rater differences about the absolute level of explaining in classrooms) and generalizability coefficients (pertaining to rater differences about the relative standing of classrooms).

Table 3 presents the generalizability and dependability coefficients for all variables for both protocols. These coefficients represent the levels of generalizability and dependability for coding using one rater and the average of the two raters. As can be seen in Table 3, generalizability and dependability were high for most variables, indicating that raters agreed highly in their coding of student explaining.

The variables with the lowest interrater consistency pertaining to correct/complete explanations in the whole-class setting. Most of the discrepancies between raters during the whole-class setting occurred when raters agreed that an explanation was provided but did not agree on whether it was correct and complete. These discrepancies occurred mostly when coding Classroom 5, in which a large number of ambiguous explanations were provided. For example, for the problem 100 + □ = 100 + 50, a student offered the explanation: “One hundred...no, the fifty will go right there because it has to mean that the 50 and the box had to be the same number.” The other rater thought that the student’s reference to “the same number” was ambiguous; if the student meant that the 50 and the box had to be the same number, the explanation could be construed as incomplete because it was not clear whether the student thought the same numbers needed to be on both sides of the equal sign or whether both sides of the equal sign needed to add up to the same number. Increasing the number of raters would not yield sufficiently reliable results for this variable. For example, even with 10 raters, the dependability coefficient for the variable “the percent of whole-class problems on which a particular student gave a correct/complete explanation” would be less than .80. Interestingly, discrepancies between raters did not emerge when they coded explanations that occurred during small-group interaction. Consequently, the dependability indices for measures that included both the whole-class and small-group contexts were quite high (see Table 3).

Video-intensive coding. Because the coding methods described above were based on a one-time viewing of videotapes...
### Table 3. Generalizability Coefficient for Different Video-Lite Variables

<table>
<thead>
<tr>
<th></th>
<th>1 Rater Generalizability Coefficient</th>
<th>1 Rater Dependability Coefficient</th>
<th>2 Raters Generalizability Coefficient</th>
<th>2 Raters Dependability Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student-Based Protocol</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of Problem Opportunities Across the Whole-Class and Small-Group Contexts in which a Student Gave an Explanation:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct/complete explanation</td>
<td>0.8297</td>
<td>0.8326</td>
<td>0.9069</td>
<td>0.9086</td>
</tr>
<tr>
<td>Any explanation</td>
<td>0.9754</td>
<td>0.9758</td>
<td>0.9875</td>
<td>0.9878</td>
</tr>
<tr>
<td><strong>Problem-Based Protocol</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of Problem Opportunities Across the Whole-Class and Small-Group Contexts in which any Student Gave an Explanation:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct/complete explanation</td>
<td>0.4901</td>
<td>0.4747</td>
<td>0.6578</td>
<td>0.6458</td>
</tr>
<tr>
<td>Any explanation</td>
<td>0.9304</td>
<td>0.9263</td>
<td>0.9640</td>
<td>0.9617</td>
</tr>
</tbody>
</table>

by raters (with no transcripts), we refer to those methods as “video-lite.” We contrast the video-lite coding with coding of the same classrooms that was performed in a previous study (Webb et al., 2009) using a much more intensive coding procedure. In the previous study, transcripts of student and teacher talk during whole-class discussions and in all recorded small-group conversations were carefully analyzed in conjunction with the videotapes. Coding of students’ explanations was based on repeated viewing of the videotapes and on repeated readings of the transcripts. The same variables were coded in the video-lite and video-intensive protocols. The only difference between the protocols is that raters watched the video only once for the video-lite protocol but were able to revisit the video and transcripts in the video-intensive protocol. For the video-intensive protocol, raters conducted multiple passes at the data, discussed their coding with other raters, and revised their coding. This protocol was much more intensive in terms of resources and time and is considered the gold standard against which to compare results for the video-lite coding.

### Results

#### Student-Based Coding

We first present results for student-based coding. The purpose of the student-based protocol was to gauge the typical level of participation of individual students in a classroom. Students who actively participate in discussions may gain through their contributions and from others’ responses to them. First, formulating and presenting ideas may promote learning as students rehearse information they already know, reorganize and clarify information to make it more coherent and understandable to others, and seek new information to make their ideas more accurate and complete (Bargh & Schul, 1980). Second, active participants may benefit from having contradictions or incompleteness of their ideas pointed out by others. Students who do not contribute their ideas will not experience these benefits. Misconceptions and gaps in understanding will remain unexposed, and opportunities to rehearse and reinforce correct understanding will be missed.

#### Video-Intensive Coding

Figure 1a presents the classroom means for student participation based on video-intensive analysis of classroom interaction using repeated viewing of the videotapes and repeated reading of the transcripts of the whole-class and small-group conversations for the eight videotaped students in each classroom. Each classroom mean is the average over the eight students in the class in the percent of problems (both whole-class and small-group) in which the student gave a correct/complete explanation. We refer to this mean as the mean-per-student percent of opportunities. For example, as can be seen in Figure 1, the six classrooms varied greatly in terms of student explanation giving. In Classroom 1, each student, on average, gave a correct/complete explanation in 22% of the problems.

#### Video-Lite Coding

Figures 1b–1e show results for video-lite coding results. Figure 1b is the video-lite analog to the video-intensive coding (the percent of problems in which a student, on average, gave a correct/complete explanation in the whole class and during small-group interaction). Next, because training raters to reliably code the accuracy and completeness of explanations may be time consuming and expensive, Figure 1c presents the results when all explanations are coded without regard to accuracy and completeness. Finally, addressing the logistical complications of coding small-group interaction, Figures 1d and 1e give the video-lite results for only the whole class (i.e., eliminating the small-group context from the coding).

As can be seen in Figure 1b, video-lite coding produced a markedly lower per-student level of explaining than video-intensive coding.
Two factors account for the difference between classroom profiles in Figures 1a and 1b. Coders were able to identify explanations in these classrooms but were not able to determine whether these explanations were complete and correct. When coders were not confident about the accuracy and completeness of explanations, they tended to not code the explanation as correct and complete. Replaying videos and rereading transcripts allowed video-intensive coders to clarify correct and complete explanations more confidently. Second, students in Classrooms 2 and 4 often gave very short explanations that video-lite coders considered as incomplete because they addressed questions that the teacher asked about one aspect of the problem, rather than the entire problem. For example, for the problem $120 + 30 = 170 - b$, when the Classroom 4 teacher asked, “Which side can I
Figure 1c presents the video-lite results when correct, complete, incorrect, incomplete, and ambiguous explanations were all counted as explanations (both whole-class and small-group contexts were coded). In contrast to Figures 1a and 1b, Classroom 6 shows a relatively high level of student explaining in Figure 1c. This is because of the large number of incomplete or inaccurate explanations in the classroom. Consider, for example, the following explanations for how to solve the problem $100 + \Box = 100 + 50$: “The fifty will go right there because it has to be the same number,” and “They have to be those because...cause it has to have the same numbers.” Such explanations were inaccurate and incomplete because the students seemed to be matching the numbers on one side of the equal sign to the same numbers on the other side of the equal sign without regard to the arithmetic operations involved or the mathematical relationship between the two sides of the number sentence. These explanations were included in Figure 1c; they were not included in Figures 1a or 1b. In Figure 1c, the incidence of student explaining in Classrooms 1 and 6 looks fairly similar. Yet, the mechanisms operating in the two classrooms were very different.

The Classroom 6 teacher tended not to push students to give correct and complete explanations. When students gave explanations that were incomplete or ambiguous, the teacher often asked only whether other students agreed or disagreed rather than probing what the student meant. In other instances, she asked questions of the student to obtain an explanation that she believed she understood, which she then revoiced more completely. Third, when students had difficulty with a problem (voicing an incorrect answer or a faulty strategy or both), she either called on another student to help or stepped in herself to direct the student to go in a certain direction without giving the original student an opportunity to develop a correct and complete explanation.

The Classroom 1 teacher, in contrast, pressed students to provide the details of their thinking, whether their answers and strategies were correct or incorrect. When a student’s work was incorrect (or incomplete), she engaged in a lengthy process of uncovering the student’s reasoning, generating a correct strategy, and making sure that the student understood and verbalized the correct strategy. As a result of this teacher’s interactions with students and the norms established in this classroom about the importance of verbalizing correct and complete explanations, nearly all conversations (both in small groups and in the whole class) developed into correct and complete explanations.

Figures 1d and 1e present the results when coding focuses on whole-class interaction exclusively. Ignoring the small-group context underestimates student explaining (compared to the video-intensive results in Figure 1a) for most classrooms. This result is not surprising, given the nature of interaction in these classrooms. For example, in Classrooms 1, 2, and 4, much student explaining occurred during small-group work. Whereas all students had an opportunity to explain their thinking during small-group conversations, only a few students were called on during the whole-class discussion of a problem. These teachers introduced a problem, asked groups to work together to solve the problem and share their thinking, then brought the whole class together for selected students to share their answers and strategies with the whole class, usually at the board. Therefore, omitting information from the small-group discussions yields a distorted view of per-student participation in these classrooms.

**Coding one day (instead of two).** Reducing the number of days of observation would decrease the amount of data to be collected and coded. Figure 2 shows the results for student participation if only one day was coded (but all other features of the coding were preserved, including correct/complete explanations, and whole-class and small-group contexts). The marks in Figure 2 represent the mean per-student explaining for each day. In Classroom 1, for example, on one day, each student, on average, provided correct/complete explanations in 22% of the problems; on the other day, the mean was 13% of the problems. The large differences in student explaining from day to day in many classrooms suggest that the particular selection of the day observed could greatly influence the results. One reason for day-to-day differences in the incidence of student explaining was the type of problems discussed. For example, in Classroom 3, the problems discussed on one day (e.g., $14/2 = [3 \times \_\_] + 1$ and $375 = \_\_ + [3 \times 10]$) were more complex and thus provided more opportunities for students to explain their thinking than problems discussed on the other day (e.g., $4000 + \_\_ + 70 + 3 = 4373$ and $308/5 = 61$ R\_\_). Another reason for day-to-day differences was the nature of the activity. In Classroom 4, for example, one day the teacher asked students to explain their strategies for solving problems such as $120 - 30 = 80 + b$. Whereas the activities on the other day concerned completing charts, for example, listing ages and the corresponding numbers of spots for the following question: “Miguel has a pet Dalmatian. Each year on the dog’s birthday, Miguel discovers three new spots on his fur. If the dog was born with four spots, how many spots will he have when he is seven?” In this activity, most student contributions focused on calculations rather than explanations of their reasoning.
Reducing (but not eliminating) the coding of small-group interaction. The results presented above show that eliminating coding of student participation during small-group work may produce distorted results. However, coding four small groups simultaneously throughout the entire class period on each occasion (as was done in this study) is logistically difficult (requiring substantial videotaping resources and careful attention to collecting adequate audio quality; see Franke et al., 2009). A more feasible approach is to code one small group (instead of four) in addition to the whole-class context. We examined two options for coding one small group: (1) coding the same group throughout small-group work, and (2) randomly selecting a different group to observe for each problem. The first option captures thorough information about a small number of students; the second option captures information for a larger number of students, but less information about each student.

Figure 3a shows the results for student participation (correct/complete explanations, whole-class and small-group contexts) when one small group is included, and that group is coded throughout the days of observation. There are four small groups in each classroom (except for Classroom 5 that had no small-group work), so each classroom has as many as four marks. Each mark represents what would happen if the same group were followed. Because there are four groups that could be followed, four marks appear in the figure. Figure 3a shows that, in some classrooms, the picture about the mean level of per-student explaining depends considerably on the particular small group selected for coding during small-group work.

Figure 3b shows the results that would be produced by focusing on one group at a time, but randomly selecting a different group on each problem. Because there were a large number of problems assigned to small groups in these classrooms, there were many different orders in which groups might be selected. Figure 3b presents 10 different randomly selected orders of coding small groups in each classroom. For example, in Classroom 1, there were 11 different problems. A random number generator was used to select which of the four groups to select to follow for each of the 11 problems. One order produced by this generating process was Group 2 for Problem 1, Group 1 for Problem 2, Group 4 for Problem 3, and so on up to Group 3 for Problem 11. This process was repeated 10 times to create 10 different orders. This approach allowed us to randomly select a different group to observe for each problem and produces more potential variability in results than keeping constant the small group observed.

Problem-Based Coding

Here we present results for problem-based coding. The purpose of the problem-based protocol was to gauge the opportunities that students had to learn by being exposed to the ideas of others. In a classroom in which student ideas are contributed for many problems, students have the opportunity to learn by listening to their classmates’ ideas. Students can compare their own knowledge to other students’ ideas, recognize and correct misconceptions and see contradictions that cause them to seek new information and help them repair imperfect mental models (Chi, 2000). Students can also learn by listening to the teacher’s responses to their classmates’ contributions. In classrooms in which few problems exhibit student participation, students will not have such opportunities to improve their understanding.

Video-Intensive Coding

Figure 4a presents the classroom profile for the percent of problems (whole-class and small-group) in which any student gave a correct/complete explanation. In all classrooms, a substantial portion of problems had a correct/complete explanation offered by a student. The much higher profile in Figure 4a than previously presented shows that correct/complete explanations emerged in many discussions of problems, even if each individual student contributed few of them.

The problem-based coding extends our understanding of classrooms from the student-based coding in several ways. Some classrooms were relatively high on both student- and problem-based coding; some classrooms were relatively low.
on both; and some classrooms were high on one but not on
the other. First, participation for students in Classroom 1 is
high for both student- and problem-based coding. From the
student-based coding (Figure 1a), we find that students in
Classroom 1 often had the opportunity to actively participate
in sharing explanations. From the problem-based coding (Fig-
ure 4a), we find that students also had frequent opportunities
to hear explanations put forth by their classmates. Thus, the
opportunities for students in this classroom represent multi-
ple ways in which students can participate and improve their
understanding.

Second, Classrooms 4, 5, and 6 are fairly low on both the
per-student participation (Figure 1a) and problem-based par-
ticipation (Figure 4a). This result shows that students in
these three classrooms did not often share their own explana-
tions nor did they have the opportunity to hear others share

FIGURE 4. Percent of problems with student explaining in each classroom (problem-based coding). a) Video-intensive coding, cor-
rect/complete explanations, whole-class and small-group; B) Video-lite coding, correct/complete explanations, whole-class and small-group;
c) Video-lite coding, any explanations, whole-class and small-group; d) Video-lite coding, correct/complete explanations, whole-class only;
e) Video-lite coding, any explanations, whole-class only.
FIGURE 5. Percent of problems with student explaining for one day only (problem-based video-lite coding, correct/complete explanations, whole-class and small-group).
Note. Each mark represents a different day.

their explanations. Thus, the opportunities for students to improve their understanding through participation in these classrooms were fairly limited.

Finally, Classroom 3 is not high in terms of per-student participation (Figure 1a) but is high on the percent of problems participation (Figure 4a). This result shows that although students in this classroom had opportunities to hear student explanations across a range of problems being discussed, participation of individual students was relatively low. One consequence of this pattern of results is that students with misconceptions or gaps in understanding may not have had opportunities to expose these problems with their thinking and have them corrected. Instead, the primary way for these students to improve their understanding was to compare their own ideas to those presented by others.

These results show that, taken together, the student- and problem-based coding provides a more comprehensive picture of student participation than either method of coding alone.

Video-Lite Coding

Figures 4b–4e present the video-lite profiles that are analogous to the video-intensive profile (Figure 4a). As was the case for student-based coding, the incidence of giving correct/complete explanations in Classrooms 2 and 4 is lower for video-lite coding than for video-intensive coding. The reasons are the same as those described earlier: confusing explanations that the video-lite coders did not count as correct/complete, and brief explanations that video-lite coders did not consider as complete. Without considering the accuracy or completeness of explanations, all classrooms show high levels of student participation (Figure 4c). When small-group work is excluded altogether from the coding (Figures 4d and 4e), the results change markedly, especially for Classroom 1, in which the incidence of explaining appears relatively low. In that classroom, much correct/complete explaining occurred in small-group problems that is not captured in the whole-class only profiles.

Coding one day (instead of two). As was the case for student-based coding, considerable day-to-day variation occurred in the picture of student explaining. As seen in Figure 5, the percent of problems with a correct/complete explanation varied markedly from one day to the next. As was the case for Figure 3, the content of problems and the nature of the activity may have accounted for some of the day-to-day variation in how many problems had correct/complete student explaining.

Reducing (but not eliminating) the coding of small-group interaction. Even if they do not have to keep track of individual students, investigators in large-scale studies may not be able to code four small groups simultaneously. Consequently, we present classroom profiles when only one group was coded per small-group problem. Figure 6a presents the results for the same group selected for each small-group problem, and Figure 6b presents the results when a different group is selected on each problem. Similar to the results to student-based coding, considerable variation in classroom results emerge depend-

FIGURE 6. Percent of problems with student explaining when only one small group is coded (problem-based video-lite coding, correct/complete explanations, whole-class and small-group). a) The same small group is coded for each problem; b) A different randomly selected group is coded for each problem.
Note. Each mark represents a different group included in the coding. Thicker marks represent similar results for different groups.
ing on the particular group chosen (Figure 6a) or the order of coding small groups in a classroom (Figure 6b).

Discussion

In large-scale studies, collecting information about classroom instruction through live observation or video recording necessitates making choices about the data to be collected and coded. This study investigated the impact of several observation design and coding choices on profiles of classroom instruction. We focused on a particular aspect of classroom instruction, student participation, as indicated by the incidence of students giving correct and complete explanations of their thinking about the strategies they used to solve mathematics problems.

The profiles of classrooms presented in this small-scale study were based on a large amount of data (e.g., multiple days of observation, a large number of students per class, multiple classroom contexts including whole-class and small-group) and distinctions between types of explanations (correct/complete versus explanations that were incorrect or incomplete). Investigators in large-scale studies may wish to (or need to) reduce the data collection resources, either by reducing the amount of data to be collected or training raters to capture subtle distinctions in the coding. We presented classroom profiles that might result by relaxing these criteria, namely, reducing the number of days of observation, reducing the number of students observed in the small-group context (or eliminating the small-group context from coding altogether), and relaxing the restriction that student explanations must be correct and complete. These choices—about whether to keep track of individual students in observations of classrooms, whether to observe classroom lessons on one or multiple days, whether to restrict attention to interaction in the whole class only and disregard small-group work, and whether to broaden the student participation dimension to include student explaining without regard to whether explanations were correct and complete—produced very different pictures regarding the level of student participation in these classrooms. That is, the characterizations of classroom interaction fluctuated depending on the coding method employed.

A large source of variation was whether to keep track of individual students in classroom observations on the resulting picture of student participation. Because of the logistical difficulty of following students, few studies do so. For example, the technical manual of the Third International Mathematics and Science Video Study (Jacobs et al., 2003) states, “While it is generally easy to distinguish the teacher’s voice from that of the students, it is not always possible to distinguish between individual student voices. Thus, there will be no attempt to track the voice of any individual student in the ongoing discourse (e.g., marking Student 1, Student 2, Student 3, etc.)” (p. 150). Our results show that observations made without following individual students cannot be used to gauge the typical experiences of students in these classrooms. For example, a classroom in which student explaining occurred for a majority of problems did not necessarily correspond to a high level of explaining on a per-student basis.

Second, the choice about whether to include or exclude student interaction that occurred away from the teacher (in our case, student–student interaction during small-group work) also had a large impact on the resulting pictures of student participation in the classrooms observed here. Because a substantial amount of student interaction occurred during small-group work without the teacher present, excluding these conversations often produced erroneous pictures of the extent of student participation. (A compromise solution we investigated—observing one small group at any one time, either the same small group or rotating across small groups during a lesson—was not a satisfactory solution because student interaction differed across small groups and the particular group chosen for observation impacted the classroom profile). The typical practice of “following the teacher” during observations, then, is likely to produce valid information about student participation only in classrooms in which students interact primarily with the teacher, and not when students spend time conversing with each other (e.g., during group work or private conversations during seatwork).

Third, the decision to use a fairly easily coded dimension of student participation (i.e., any explanation that a student provided without regard to accuracy and completeness)—rather than a more nuanced dimension of student participation that was found to predict student achievement in previous studies (i.e., explanations that were correct and complete)—also produced erroneous results in some classrooms. The stripped-down version of coding student participation that does not consider the quality of student explanations may be easier to implement but it did not always produce an accurate description of a classroom.

This study, then, found that stripping away elements of observation and coding (e.g., coding classroom interaction without regard to contributions of individual students, capturing only whole-class instruction and not other contexts that afford opportunities for student participation, using less nuanced rather than more nuanced dimensions of student participation) did not fully characterize, or even misrepresented, the level and nature of student participation in many classrooms. For purposes of providing feedback to teachers to improve instructional quality, information about per-student participation across the range of classroom contexts, as well as the extent to which students are providing explanations that are accurate and complete, is important and useful. For example, teachers who learn that students, on average, participate in few classroom conversations may make efforts to include more students in discussions or to provide more opportunities for students to share their thinking. Teachers who provide opportunities for students to participate but who learn that the incidence of correct and complete explaining is low may make more effort to identify, discuss, and rectify student misconceptions and gaps in understanding.

Although the emphasis in this paper is on providing information that teachers could use to improve the quality of their instruction, we recognize that such information could also be used to make comparisons among teachers. Although not explored in depth here, the results of this study suggest that the impact of observational design choices also applies to teacher comparisons. As is evident in the findings reported and discussed here, which teachers may be identified as “higher” than others in terms of student participation may vary markedly according to the observation and coding choices made. In conclusion, for a variety of purposes of collecting observational data in classrooms, the fluctuation of results across the different observation and coding choices examined in this study point to the importance of carefully considering methods used to characterize instructional opportunities in classrooms.
Acknowledgments
This work was supported in part by the Spencer Foundation; the National Science Foundation (MDR-8550236 and MDR-8955346); the Academic Senate on Research, Los Angeles Division, University of California; the Diversity in Mathematics Education Center for Learning and Teaching (DIME); and by Grant No. 1093264 from WestED to the Center for the Study of Evaluation/CRESST. Funding to WestED was provided by Grant No. ESI-0119790 from the National Science Foundation. Funding to DIME was provided by Grant No. ESI-0119732 from the National Science Foundation. We thank Melissa Cook, Megan Franke, and Angela Chan Turrou for their valuable feedback. However, all opinions in this article are those of the authors.

Notes
1 In the small-group context, for a particular problem, if any student within any small group offered an explanation, it was counted as one instance of an explanation for this problem (regardless of the number of students or groups who provided an explanation on that small-group problem).
2 The sample size reported in the study was larger than in this study. The previous study included students recorded using audio setups without videotaping as well as students appearing on videotape. This study includes only students who appeared on videotape.
3 The conclusions drawn from coding one group were similar for all explanations (regardless of accuracy and completeness) and are not presented here. In addition, conclusions drawn from randomly selecting two groups instead of one group were also similar and are not presented here.

References


