

# The Use of Qualitative Methods in Large-Scale Evaluation: Improving the Quality of the Evaluation and the Meaningfulness of the Findings

JULIE SLAYTON

*Los Angeles Unified School District*

LORENA LLOSA

*New York University*

*In light of the current debate over the meaning of scientifically based research, we argue that qualitative methods should be an essential part of large-scale program evaluations if program effectiveness is to be determined and understood. This article chronicles the challenges involved in incorporating qualitative methods into the large-scale evaluation of the Waterford Early Reading Program and the decisions made in light of those challenges. We demonstrate that, in spite of the challenges, there are significant benefits associated with using qualitative methods on a large scale. More specifically, by using qualitative methods, we were able to improve the evaluation and generate findings that were meaningful and useful to stakeholders. It is our hope that our experiences can be used by others to inform their work as they seek to incorporate qualitative methods into their large-scale evaluations.*

## INTRODUCTION

The No Child Left Behind Act of 2001 (NCLB) has prompted educational researchers to conduct *scientifically based research* to inform instructional practices and support the adoption of instructional programs. It has also prompted educational researchers to debate what is meant by the term *scientifically based research* and what methodologies are legitimate, appropriate, and in line with such research efforts (Eisenhart & Towne, 2003). The most prominent interpretation of the legislation is that scientifically based research is limited to experimental or quasi-experimental designs (Slavin, 2002) and quantitative methodologies. Others argue that scientifically based

research should offer opportunities to engage in a variety of designs and methods, allowing the research question to drive the design and methods adopted (Burkhardt & Schoenfeld, 2003; Chatterji, 2004; Erickson & Gutierrez, 2002; Feurer, Towne, & Shavelson, 2002).

In fact, Maxwell (2004) made a strong case for not limiting scientifically based research to experimental and other quantitative methods. He argued that “qualitative research is a rigorous means of investigating causality” (p. 3) and should be an integral part of scientifically based research in education. We agree with Maxwell—that to understand study outcomes, quantitative methods are insufficient. In program evaluation, for example, when no evidence of effectiveness is found using quantitative methods, qualitative methods can explain why the program was not successful. These explanations might include issues such as the following: The materials might be too difficult to implement, the teacher might not be implementing the program appropriately, the program might not be engaging and the students do not pay attention, or the professional development is insufficient to support the program implementation. When a program is determined to be effective, qualitative methods can confirm that it is actually the program that is responsible for the effect. Qualitative methods pose the questions of how and why a program does or does not cause the intended effect and not simply whether a program causes the intended effect. Therefore, we believe that qualitative methods should be an essential part of large-scale program evaluations if program effectiveness is to be determined and understood.

The purpose of this article is to document the challenges and benefits associated with a large-scale evaluation that incorporated qualitative methods. We used qualitative methods in order to gain a “greater program understanding and more explanatory power, specifically about *why and how* certain outcomes were attained or not” (Greene, 1998, 141). Below, we describe the evaluation. We chronicle the challenges we faced and the decisions we made in light of those challenges. Finally, we document the benefits of using qualitative methods to inform the evaluation itself and generate findings that were meaningful and useful to stakeholders. We hope that our experiences can inform others who are interested in or actively attempting to incorporate qualitative methods into their large-scale evaluations.

## BACKGROUND

In 2001, the Program Evaluation and Research Branch was asked to evaluate a computer-based reading program adopted by the Los Angeles Unified School District (LAUSD) to supplement an existing reading program in kindergarten and first-grade classrooms. The district had adopted the Waterford Early Reading Program (Waterford program) in 244 schools and 2,235 classrooms, to be used with 81,000 students in kindergarten and first

grade. Our branch was already responsible for evaluating the implementation of the district's primary reading program, Open Court Reading (Open Court). Now, we were being asked to examine the extent to which this new program, Waterford, impacted student achievement and teacher practice when it was used as a supplement to the Open Court Reading program. In kindergarten, each child was expected to spend 15 minutes a day working individually at one of the three Waterford computers placed in the classroom for that purpose. In first grade, children were expected to spend 30 minutes a day at one of the three computers with the courseware. The program's courseware could be personalized for each child's learning pace and reading level, and the program could provide teachers up-to-date information on their students' progress and needs as they moved through the various reading levels. In addition, children had a series of books and videotapes to take home, thus giving them further exposure to the material and giving parents the opportunity to read with their children.

## RESEARCH QUESTIONS AND STUDY DESIGN

Over the course of the 3-year evaluation period, we attempted to address a range of research questions. The two overarching research questions we addressed were as follows:

1. Does the Waterford Early Reading Program, as implemented in the LAUSD, have an effect on student reading achievement?
2. To what extent is the Waterford courseware being implemented?

To address the research questions, we had no choice but to adopt a quasi-experimental design. One of the realities of program adoptions in school districts is that these decisions are driven by district policy and not by evaluators interested in determining the effect of a particular program or intervention. As a result, evaluators, by and large, cannot control how students, teachers, and schools are assigned to receive a given program or intervention. This reality makes experimental design nearly impossible and forces evaluators to use quasi-experimental design at best, and more often, one-group designs. In this case, we were fortunate that the district did not choose to implement the program districtwide, but only in a subset of schools meeting a certain set of criteria. As a result, we were able to identify a comparable group among those schools that were not receiving the Waterford program and use a quasi-experimental design. The sample of our study consisted of 2,000 students in 100 kindergarten and 100 first-grade classrooms. Thus, in the 50 kindergarten and 50 first-grade treatment classrooms, the Waterford program and Open Court were being

implemented. In the other 50 kindergarten and 50 first-grade classrooms, students were only exposed to Open Court.

Our desire to gain a deeper understanding of the program implementation and effect caused us to seek rich narrative data in addition to the test score data necessary to compare the outcomes between the treatment and comparison groups. Therefore, the data collection included extended observations in all 200 classrooms requiring narrative scripting. We also conducted interviews with each teacher and a set of administrators to elicit information regarding teacher use and knowledge of both the Waterford program and Open Court. In addition, we collected quantitative data that included usage data from the Waterford program, individual student test data from our administration of the Woodcock Reading Mastery–Revised (Woodcock, 1998), and the administration of the statewide reading standardized tests.

As will be discussed below, both the requirement to have such a large sample in order to generalize our findings to our very large and diverse district and our desire to incorporate rich narrative data created significant challenges to the overall management of the evaluation in all its stages, particularly qualitative data collection and analysis.

## CHALLENGES OF INCORPORATING QUALITATIVE DATA INTO LARGE-SCALE EVALUATION

We were confronted by a number of difficulties in incorporating qualitative data into our evaluation. We will present them in two separate sections. The first focuses on issues related to planning and executing the data collection, and the second focuses on issues related to analysis. In terms of data collection, challenges included overcoming significant time limitations in getting into the field, sufficiently staffing the project, ensuring the quality of the data being collected, guaranteeing confidentiality and anonymity, and managing multiple rounds of data collection. Below we describe each of these obstacles and how we handled them.

### DATA COLLECTION

#### *Time Limitations*

Time was a significant issue for conducting classroom observations because we wanted to gather information related to initial program implementation and changes in practice/implementation over the course of the school year. We also needed to be able to compare instructional practices in relation to Open Court in treatment and comparison classrooms. Thus, we decided to

observe each classroom for 2 days during reading/language arts in both the fall and spring. We needed to conduct our observations within a narrow window of time. Thus, we sought to observe between the 8th and 12th weeks of the school year and then again between the 8th and 4th weeks before the end of the school year. In addition, we had to account for the fact that we operate on a year-round calendar. Some of our schools operate on a traditional calendar, with all students attending from September through June. Others operate with multiple tracks of students from July to June. For example, on a three-track calendar school, for Track A, which begins in September, weeks 8–12 fall in October and November, whereas Track B begins in July, and the 8th week of school is at the end of August. Students then go off track for 2 months, and weeks 9–12 are not until November. Track C, on the other hand, begins in July, and weeks 8–12 fall in August and September. Consequently, we had to adjust to the multiple track schedules, which meant that we had multiple data collection schedules.

### *Organizational Challenges*

In addition to the challenges posed by time constraints, the organizational challenges we encountered to ensure that observations took place within our designated data collection windows were formidable. Although the Program Evaluation and Research Branch staff of 30 researchers is large when compared with other school district evaluation offices, no more than one or two researchers are available to direct an evaluation at any time, and many are responsible for more than one project. As a result, we were dependent on a team of 20–30 individuals who worked for us on a part-time basis to do most of the data collection. Because of the nature of the job—data collection was cyclical and the hours were inconsistent—we were constantly faced with having to recruit and train new data collectors. Because of the substantial focus on a very detailed narrative type of qualitative data collection, which required someone who had the intrinsic strengths of a researcher—inquisitive, detail oriented, organized, focused, flexible, hard working—our preference was to hire graduate students in master’s or Ph.D. programs. Although we were surrounded by several major universities with graduate programs, we were competing for these students not only with the graduate schools that employed them as graduate student researchers and teaching assistants but also with other research institutions in the area. With the universities, graduate student researchers received an hourly salary, and at least some part of their tuition was paid for. The students also received health insurance from their schools. Although we could pay a similar hourly rate, we could not provide any benefits to temporary employees. Similarly, we were unable to pay an hourly rate that was competitive with

the research organizations in the area. Consequently, our research team was composed of some graduate students and recent college graduates.

In addition to the difficulties we faced hiring staff with the requisite qualifications, our reliance on this large number of transient employees created a host of challenges to data-collection activities. Each round, we were faced with having to replace data collectors because of attrition. In addition, we spent significant amounts of time supervising and mentoring data collectors: talking on the phone, communicating with them via e-mail at least once a week regarding their progress, meeting with them in person, reviewing their work, and answering questions. Dedicating time in this way was critical for us, though, because we depended on data collectors to get the data we needed to conduct an effective and comprehensive evaluation.

### *Protocol Development*

Prior to entering the field, we discussed the types of data that we would need to collect in order to answer our research questions. We knew that we wanted as much information as we could gather about each classroom in the context of reading/language arts instruction, particularly student use of the Waterford courseware, teacher pedagogy, and student behavior. Others have used checklists, responses to directed questions, summaries of classroom activities, and ratings (Aschbacher, 1999; Datnow & Yonezawa, 2004; Noga, 2002; Piburn & Sawada, 2001; Saxe, Gearhart, & Seltzer, 1999), but we decided that they were limited for a variety of reasons. Checklists required predetermined categories, and we did not know enough about what we should expect to see in the classroom. In addition, checklists required that data collectors be knowledgeable about the courseware, the Open Court curriculum, and reading/language arts instructional practice in order to understand the established categories. It is difficult to establish consistency with checklists and summaries because each data collector is coding the data while collecting them, which requires a great deal of content expertise. More important, because data reduction takes place at the point of data collection when using a checklist or summary, the data can be used for quantitative analysis but not for in-depth qualitative analysis. Consequently, we knew that we did not want to use a checklist or a summary-type protocol. Instead, we turned to open-ended narrative-based protocols to overcome the limitations of checklists and summaries.

We felt that narrative-based observation protocols would remove the responsibility for making coding decisions from the data collector. His or her role would be to record the observation with sufficient detail to allow us to code and analyze the data. Additionally, a narrative-based observation would capture data in a form that could be used for both qualitative and

quantitative analysis. There would be sufficient information to quantify those items that would be captured in a checklist and to provide the details and nuances of teacher pedagogy and student behavior that would be essential to understanding the reasons behind the evaluation outcomes and that could inform future practice and program implementation. We planned on creating narrative-based protocols that would require data collectors to write field notes. We defined field notes as a written narrative describing, in concrete terms and in great detail, the duration and nature of the activities and interactions observed. Classroom data collectors would become the “eyes and ears” of the project, and their notes would describe the overall context in which reading/language arts instruction took place.

We designed one narrative-based observation protocol to capture data about the use of the Waterford courseware and a separate narrative-based observation protocol to collect data regarding teacher practice and student behavior in the context of Open Court instruction. The rationale for a protocol capturing teacher practice and student behavior was that we would be able to compare our treatment and comparison classrooms and thus separate what students were learning during Open Court instruction from what they were learning from the Waterford courseware.

The Waterford computer log was designed to capture information from three students simultaneously as they each worked individually on the courseware at a computer station. We needed the protocol to record the students’ names, the time they began using the courseware, and the time they completed their turn, in addition to a detailed narrative about the students’ behavior while using the courseware. The name and time information would be necessary to relate students’ time spent on the computer with their test scores. The students’ names and the narrative would document the students’ level of engagement with the courseware, anything that might impact their turn (i.e., computer problems and interruptions), and the extent to which they fully used the program. We could also use this data to examine whether there was a relationship between level of engagement and test scores. Figure 1 shows an excerpt from our Waterford computer log for one student. The full protocol contains room for three students per page.

The classroom observation protocol was designed to document classroom interactions during reading/language arts. The prompt provided to data collectors is presented in Figure 2. Pages of lined paper follow the prompt.

In addition to the observation protocols, we felt that we would need a supplemental protocol for data collectors to gather some initial information regarding the classes they were to observe. Thus, we crafted a one-page initial interview intended to take no more than 5 minutes. In treatment classrooms, we wanted to know about teacher training on the Waterford courseware, computer installation, whether all the equipment was there

Teacher		
Code:	Location Code:	Date:
Grade:	Observer:	
COMPUTER STATION 1: Name: _____ Time In: _____ Time Out: _____		
_____		
_____		
_____		
_____		

Figure 1. Waterford Computer Log

and working, and whether teachers were using materials intended to support courseware use in the students' homes. For both treatment and comparison classrooms, we also gathered general information about the classroom, such as students' English language development levels, the teacher's lesson goals and objectives for the day, and the Open Court unit and lesson that they would be teaching. The challenge in developing this protocol was to make it comprehensive, yet brief enough to ensure that teachers would give their time to answer the questions before class started for the day.

Finally, we wanted to ensure that our data collectors would have the opportunity to record any information or data gathered during their observation that did not belong on any of the other observation protocols. This additional protocol would also be a place for data collectors to document their experiences, biases, and likes and dislikes of a classroom observation experience. It would allow the data collector to intentionally place any subjective comments they had regarding their observation so that they could avoid expressing these comments within the context of the objective field notes taken during the observations. Thus, we designed a reflective note protocol that asked the data collector to reflect on the two days of observation.

In addition to classroom practice, we wanted to capture information related to teachers' experiences using the Waterford program and Open

<p>In a narrative, describe the activities taking place in the classroom. Focus on teacher-to-student(s) and student-to-student interaction, as well as activities of individual students. A narrative requires that you directly quote the teacher as he or she asks questions, models reading strategies, presents material, and works with students in a whole group, small groups, or individually. Additionally, you should document student engagement in response to the teacher and individual or group work and nonwork activities.</p>
--

Figure 2. Classroom Observation Protocol Prompt

Court during reading/language arts instruction. We designed one interview protocol for treatment teachers and one for comparison teachers. Each interview took approximately 30 minutes to administer. We asked both comparison and treatment teachers questions regarding the Open Court training and coaching that they received, their perceptions about the effectiveness of Open Court, and their implementation of Open Court. For treatment teachers, we also asked an additional set of questions specifically geared toward the Waterford program implementation. The topics of the interviews included training, frequency of use of the courseware and supplementary materials, the way that the teacher used Waterford during Open Court instruction, whether the courseware fully engaged students at all times, the features of the program, and perceptions of the effectiveness of the courseware.

### *Training the Data Collector*

Given the demanding nature of our protocols, it was clear to us that we would have to extensively train our data collectors to ensure the quality of the data collected. Moreover, because we would have to observe 200 classrooms within a period of 2–3 months, our challenge was to train a sufficient number of data collectors to gather high-quality, detailed data. Training thus became a critical component of the evaluation. This was also an extremely time-consuming activity; we had to dedicate 3 to 4 full days to training. Because we had a fall and a spring data collection and multiple tracks for each round, we had to conduct four observation training sessions per year.

During the first day of training, an overview of the evaluation project was presented, and the data collectors were introduced to the protocols. Additionally, information was provided regarding the procedures for contacting schools and scheduling observations. The second day consisted of a mock observation. Data collectors, accompanied by one of us, observed a class for an hour and were then given feedback on their notes. On the third day, the data collectors worked on several activities to reinforce the strategies for documenting their observation. In addition, we shared with data collectors some of the coding and analysis that would be done using the data that they collected. In the spring, we included interview training because data collectors were also responsible for interviewing teachers at each school in which they conducted an observation. For spring observations, we required all returning data collectors to participate in a 1-day retraining session during which they were given the opportunity to self-critique their observation notes from the fall.

After training, we gave each data collector one classroom to observe. We required that they return to us so that we could review their field notes after

completion of the first observation. This allowed us to monitor and provide feedback to ensure that the data we were collecting met our expectations. In some cases, this led to additional meetings with the data collectors. We learned that this level of attention to training data collectors was critical to our ability to collect high-quality, detailed observation data that focused on the interactional dynamics of the classroom situation—that is, the “immediate environment of learning” (Erickson, 1988)—and link these dynamics to student achievement.

### *Confidentiality, Anonymity, and Building Teacher and Administrator Trust*

We were presented with a series of obstacles as we entered schools to observe classrooms. At a time when schools found themselves the subject of a multitude of evaluations from grant providers, the state, federal programs, and our district office, many principals would have preferred to exclude their schools from our evaluations because of the burden that they believed our presence placed on their teachers. Even though teachers and principals were required to participate, we still had to persuade them to allow us into their classrooms and schools.

In the correspondence we sent out, we informed teachers and administrators that our sole purpose in their schools and classrooms was to observe program implementation and not to evaluate a particular teacher or school. However, it was often difficult for teachers and administrators to understand the distinction that we were making. Although we explicitly explained that our purpose in the schools and classrooms was separate from any type of teacher evaluation, and that we did not use teacher or school names when reporting findings, teachers often felt threatened by our presence in their classrooms. The impact of this tension on our ability to collect data was that we often found ourselves talking to teachers and principals to explain our purpose in their schools and assuage their reservations and concerns. This activity consumed a great deal of our time, and in some cases, it required that we go to the schools to conduct observations if a data collector was unable to gain access. It was important that every classroom selected for the study actually participate because allowing teachers and administrators who did not wish to participate in the study to opt out compromised the generalizability of the study and, more broadly, the branch’s ability to conduct any future evaluation of districtwide program implementations.

An obstacle that we needed to overcome in the Waterford program evaluation was explaining to teachers in comparison classrooms why they had been selected to participate in an evaluation of a program that they did not have. In fact, in some cases, administrators voiced concern that

participation as a comparison school might exclude them from receiving the program in the future. Again, we spent time during each round of data collection talking to principals and teachers to assure them that their participation would in no way impact whether their school would receive the Waterford program in the future, and that they were selected as comparison schools so that we might determine whether the Waterford program was impacting student achievement above and beyond the impact of Open Court.

### CHALLENGES IN QUALITATIVE DATA ANALYSIS

A significant challenge presented by the size and scope of the Waterford Evaluation design was the use of traditional qualitative methodologies to collect and analyze data about program implementation within a large-scale multisite study. Because the Waterford program is a supplemental program, we needed to understand both the extent to which it was being implemented as a reading program independent of anything else happening in the classroom and how it was implemented in relation to Open Court. Moreover, because the design of the study was quasi-experimental, we needed to be able to assess the quality of implementation of Open Court in the comparison classrooms as well. The primary difficulties related to analysis were that we had thousands of pages of qualitative data to analyze—detailed narrative Waterford program use notes collected using the computer log protocol for 100 classrooms over 4 days and detailed narrative Open Court notes collected using the classroom observation protocol for 200 classrooms over 4 days—and all of our analysis had to be completed in less than 5 months so that we could provide a report to the Board of Education by December.

Within the analysis, many challenges presented themselves. The qualitative data analysis had to focus on a number of separate but related sets of data. In treatment classrooms, the analytic focus included (1) teacher pedagogy and its relationship to the implementation of the Waterford program and Open Court and (2) student behavior in relation to the Waterford program and teacher pedagogy. In comparison classrooms, the analytic focus was limited to teacher pedagogy in relation to the implementation of Open Court and student behavior in relation to teacher pedagogy. To determine the extent to which the Waterford program was responsible for increased reading ability, we had to be able to disentangle what students were learning through their usage of the Waterford program from what they were learning from the teacher and Open Court.

Thus, we adopted a combination of strategies to accomplish our goal. To be able to analyze the tremendous amount of data collected, we sought to

create a research team similar to those typically found at graduate schools employing graduate student researchers led by a primary investigator. Thus, we hired six graduate students to work closely with us on data coding and initial analysis of the Waterford program data and the data on teacher-directed Open Court instruction. Each member of the team was responsible for coding a grade level and track for both fall and spring rounds of data collection and then for both treatment and comparison classrooms. Our general approach to qualitative data analysis was to use inductive analysis (Patton, 2002). As we moved forward through the coding process, we met weekly to identify themes and categories, after which each person went back and reexamined his or her data to verify the presence or absence of a given theme or category and to ensure consistency in the process. We began with specific observations of teacher and student behavior and built toward general patterns of behavior.

#### WATERFORD PROGRAM DATA CODING AND ANALYSIS

For the Waterford program data, the team of analysts examined notes for each student who used the courseware in our 100 treatment classrooms. The notes provided sufficient detail for the analyst to determine the extent to which the student was engaged. As can be seen in the following examples, observation notes captured a range of behaviors by students during their use of the Waterford courseware. We were generally able to tell if a student was highly engaged, following the directions of the courseware (Example 1); distracted by other students (Example 2); or interrupting other students working on other computers or in large group activities with the teacher (Example 3).

##### Example 1

Maria puts earphones on. Very engaged. Works diligently. Takes the microphone and speaks into it. Prints something out. There's a chicken with eggs on the screen. She puts the printout next to her, continues to work. Prints again. Letter F, D, R on the screen. Follows the "X" with her finger on the screen. She's making letter sounds. Is very engaged!

##### Example 2

She is talking to Joy. Fighting with Joy. Made up. Talking to Joy about what's on her screen. Headphones off. Not watching. Recess. Participating in lesson. Not looking at computer. Watching Joy's screen. Talking to Joy re: her screen. Talking but not about lesson. Waiting on computer screen.

### Example 3

Jacinta watches photos of city and country. Distracted by Nikki's singing. The City Mouse and the Country Mouse, long story. She pays close attention to screen the whole time. Now follow-up activities for story. She says "Games!" Word puzzle, parrot picture underneath. Keeps saying "Awesome" as picture is revealed. Distracts Nikki and another student. Likes the games and gets excited. Sings aloud. Distracts Alan and another girl. "ed" on screen. Jacinta is singing to a song—loud. Teacher says to her, "Shhh." Jacinta is very into songs. Aide and Teacher tell her "Shhh." "Whale, wh, wh," song. Sings aloud. Jacinta ignores Nikki—sings and sings.

The detail present in our data allowed us to create a rubric that defined different levels of student engagement. The rubric was a 4-point scale ranging from *fully engaged* to *completely off-task*. We then were able to apply that rubric to make determinations about the level of engagement, the reasons behind that level, and other factors that affected students' turns at the computer (interruptions, computer problems, Open Court instruction).

While we were working on the individual student level analysis, the 4 days of data for each classroom allowed for classroom-level themes to emerge, such as computer problems and how they affected the use of the courseware, the relationship between student level of engagement during Waterford program usage and during Open Court instruction, and English learners and their engagement with the courseware as compared with their engagement during large group instruction and English-only students.

#### OPEN COURT DATA CODING AND ANALYSIS

For the Open Court data, the team examined notes for each reading/language arts activity. The notes provided sufficient detail for the analyst to examine in greater depth the themes and categories identified during the initial coding. Below is an excerpt from field notes taken during a kindergarten "Preparing to Read" lesson:

T: J is the magic letter. We are going to learn the poem for the letter J. All right. Her name, her name is Jenny. I have a friend named Jenny!

Ss: Jennifer!

The teacher demonstrates juggling.

T: I tried to juggle at home with oranges. But you know what, they got squashed.

The teacher introduces the /j/ sound.

T: Softly /j/ . . . /j/ . . . /j/ . . . Let's see if I get this sound right.

The teacher reads the poem to the students and then turns on a tape of a song that says "the J sound goes /j/ /j/ /j/."

She now introduces the *Jj* card pictures. She holds up cards with the following words: jar, judge, juice, jellyfish, and Jello. The teacher describes each word and gives a Spanish translation.

The teacher now passes out *Jj* letter cards.

T: Is it a fan?

Ss: No.

T: Do you put it in your nose? I'm gonna say some words and if they begin with the /j/ sound, you lift it.

T: Green.

Some students lift their cards.

T: Did it say jreen?

The teacher goes through additional words, repeating the same pattern.

T: Now for my magic bag. I'm thinking of something that I drink in the morning.

S: Orange juice.

T: Yes, it can be apple juice, grapefruit juice, or prune juice because one day I'll be old and need prune juice.

The teacher is now asking questions asking students to identify j words she pulls from her bag. She takes out a box of Jello. She spells out Jello in a humorous way.

Ss: Gelatina!

T: I'm going to put the container like this so you can see the letter j in Jello.

T: Jelly beans and the company that makes them are called Jolly Ranchers. Jolly means happy. He is a happy man.

The teacher takes out a jump rope.

T: This is a jump rope. I trained with the best. The rope needs to jump. I think I know what to do. But you have to help. Every time the rope touches the ground you have to go /j/.

The teacher starts jumping rope. Students say **jah**.

T: Not **jah**. Its /j/, softly. I'm gonna do it one more time, but my heart . . . I think that's about it. Go back to your usual spots on the rug.

The teacher approaches the students individually. Each student says j.

T: I could have had a cardiac arrest today, but it would have been worth it. You learned it!

The observation notes, like those above, were examined in relation to the Open Court teachers' manuals and relevant research literature to assess the quality of the implementation and teacher pedagogy. For example, for the data presented above, the teacher's manual instructs the teacher to introduce the Jj sound as follows:

- Display the *Jj Alphabet Card* and say the sound of the letter, /j/. Show the picture for the /j/ sound and teach the short poem for /j/:

Jenny and Jackson like to have fun.

They play jacks, jump rope, and juggle in the sun.

Each time they jump, their feet hit the ground.

/j/ /j/ /j/ /j/ /j/ is the jumping-rope sound.

- Repeat the poem, emphasizing the initial /j/.

Next, the lesson directs the teacher to have the students listen for the initial /j/ sound:

- Hold up and name each of these **Picture Cards**: jam, jar, judge, jeans, juice, and jellyfish. Ask the children to listen for the /j/ sound at the beginning of the words.
- Give each child a *Jj Letter Card*.
- Have the children hold up their *Jj* cards each time they hear the /j/ sound at the beginning of the words. Try these words:

---

Green	jail	Jeans	Gail	Jill
Jake	Jam	Jim	gas	

---

In the next section, the teacher was directed to play the “I’m Thinking of Something that Starts with \_\_\_” game. According to the teacher’s manual, the teacher should:

- Play the “I’m Thinking of Something that Starts with \_\_\_\_\_” game, using words that begin with /j/. Choose objects that are outside of the room but give the children some clues to what you are thinking of. You might try the following objects and clues:

Something you drink in the morning (*juice*)

Something you put on your toast (*jam* or *jelly*)

The sound bells make (*jingle*)

- If you have children in your class whose names begin with /j/, you might want to use their names in the game.

After analyzing a significant portion of the notes and determining themes and categories by comparing the notes with the teacher’s manual, we created a rubric to assess the quality of pedagogy combined with the fidelity of program implementation across the 200 classrooms of data. A 5-point scale was constructed to reflect high- to low-quality pedagogy and program fidelity (Slayton & Llosa, 2002). Our definition of high-quality pedagogy was developed after an extensive review of the research literature on reading and comprehension instruction (Adams, 1990; Duffy, Roehler, & Herrmann, 1988; Ehri, 1992; Yopp, 1992; Pearson, 1984; Pressley et al., 1992). In this way, data from all four days of observation in the 200 classrooms were examined by grade and by track for overall findings related to implementation and pedagogy.

The types of analyses we undertook were labor- and time-intensive. However, they were intended to provide decision makers with concrete information to understand the *quality* of implementation and pedagogy and student engagement. In addition, we used the rubric scores and the ratings of student engagement as variables in our quantitative analyses of program effect (Slayton & Llosa, 2002). The connections between achievement and pedagogy or implementation would allow us to examine the role of the teacher in impacting program implementation and student achievement.

## BENEFITS OF QUALITATIVE DATA AND ANALYSIS

Our choice to incorporate rich narrative observation data from the outset of the evaluation had a number of benefits. By observing teaching practice and program implementation from the beginning, we were able to gain a

thorough understanding of the context in which the program was being implemented, including the quality of the pedagogy provided by teachers, the range of student behavior while using the program, and the overlap between Open Court and the Waterford program during reading/language arts instructional time. One benefit was that we were able to eliminate potential explanations for differences between the treatment and comparison groups and make thoughtful changes to the evaluation plan to redistribute resources and narrow the focus of the evaluation to those elements that were most likely to be responsible for program effect. A second benefit of using these rich narrative observation data was that they added significantly to the usefulness of the findings for policy-making purposes.

#### TO IMPROVE THE EVALUATION

As we discussed earlier, when the evaluation began, one of our concerns was the supplementary nature of the Waterford program. To determine whether the Waterford program was responsible for any benefits beyond those resulting from the primary instruction, we first had to establish that the quality of pedagogy was comparable in treatment and comparison classrooms. Thus, in the first year, we observed treatment and comparison classrooms for two days in the fall and the spring. We found no differences in the quality of pedagogy in kindergarten and first-grade treatment and comparison classrooms (Slayton & Llosa, 2002). This assured us that any differences in achievement that we might find would be the result of the use of the Waterford program. This finding also allowed us to eliminate quality of pedagogy as a primary focus for future data collection and analysis.

We used our experience in the first year to make changes to our data collection and analysis in the second year of the evaluation. We turned our attention to a more careful examination of the treatment classrooms in terms of students' level of engagement and the interaction between the primary reading program and the Waterford program. To address these issues more directly, we revised our Waterford computer log (see Figure 3).

In addition to the space provided for narrative notes, we included the ratings of engagement that we developed in the first year. These ratings ranged from fully engaged to off-task. We also provided data collectors with a space to identify the type of activity—OC Green Section, OC Red Section, OC Blue Section, or other—under way in the classroom during each student's turn on the Waterford program. These activities refer to different components of the Open Court program. The Open Court 2000 Teacher Edition is divided into three separate components: Preparing to Read, Reading and Responding, and Integrating the Curriculum. Each section is color-coded. The Preparing to Read component is green and is often referred to as the Green Section. This section focuses on decoding and

Teacher Code:	Location Code:	Date:
Grade:	Observer:	
COMPUTER STATION 1:      Name: _____      Time In: _____ Time Out: _____		
_____		
_____		
_____		
_____		
<input type="checkbox"/> Fully engaged <input type="checkbox"/> Minor distractions <input type="checkbox"/> Distracted <input type="checkbox"/> Off-task		
Classroom activities: <input type="checkbox"/> OC Green Section <input type="checkbox"/> OC Red Section <input type="checkbox"/> OC Blue Section <input type="checkbox"/> Other: _____		

Figure 3. Revised Waterford Computer Log

fluency. The Reading and Responding component is color-coded red and is referred to as the Red Section. This component focuses on reading comprehension instruction. The Integrating the Curriculum component is color-coded blue and is referred to as the Blue Section. This component focuses on language arts and writing.

These changes to the protocol also required changes in our data-collector training. These changes took two forms. The first was to train our data collectors to be able to characterize each student's level of engagement as fully engaged, minor distractions, distracted, or off-task. This understanding would allow them to check the appropriate box while conducting the observation. In addition, we had to provide additional training for the narrative note-taking. We focused more attention on the process of documenting the specific interaction between the student and the computer. In other words, we trained our data collectors not to characterize student behavior as engaged or on-task, but instead to carefully describe the specific actions in which students were engaged while sitting in front of the computer. For example, a student might trace the letters on the computer screen, sing along with a song, stick a pencil in the disk drive, point to a neighbor's screen, or look away from the computer. In the first year, while data collectors attempted to capture these behaviors, they would sometimes summarize the student's behavior as being engaged or being off-task (see Example 1 presented previously). These overgeneralizations of student behavior as engaged or off-task made it difficult to ensure that our interpretations of the students' levels of engagement were reliable across data collectors during the first-year data analysis phase. At the time, we also did not want the data collector coding student engagement because we did not yet have a common set of definitions for engagement and had no realistic understanding of what

student engagement with the program would look like until we had observed it. Instead, we wanted narrative descriptive notes to allow the analysis team to develop a coding scheme that would be applied consistently across all observations. In the second year, on the other hand, we did have a very clear understanding of the range of student behaviors while interacting with the program. Therefore, we felt comfortable adding the ratings to the protocol.

Our findings in the second year also shaped our third-year data-collection plans. In the second year of data collection and analysis, we also found no differences in the quality of pedagogy in treatment and comparison classrooms. Consequently, we decided not to collect classroom observation data from comparison classrooms in the third year. Additionally, in the first two years, we found no changes in pedagogy from the fall to spring rounds, so we also made the decision not to conduct two rounds of classroom observations, but to only observe treatment classrooms in the spring. Because we did not want to lose representativeness in the teacher use of the program, we did one 3-day observation instead of the two 2-day observations that we had conducted during the previous 2 years.

#### TO PROVIDE MEANINGFUL FINDINGS AND INFORM POLICY

In addition to benefiting the evaluation process, the incorporation of rich narrative qualitative data added significantly to our ability to interpret the findings. Through the quantitative analyses, we found no differences in reading achievement between students exposed to the Waterford program and those who were not. This was true in the first and second years of the evaluation (Hansen, Llosa, & Slayton, 2004; Slayton & Llosa, 2002). The narrative observation data allowed us to explore the possible explanations for our findings. We were able to consider whether varying qualities of teacher pedagogy or implementation of Open Court in treatment and comparison classrooms were responsible for our results. We also considered whether the varying levels of student engagement both during primary instruction and during Waterford program use explained the findings. Finally, we examined the possibility that the reason why we found no differences was that the Waterford program was being used to supplant rather than supplement primary Open Court instructional time that focused on developing students' knowledge of sounds and letters, decoding, and fluency. We will discuss each of these in turn and explain how our findings allowed us to make specific recommendations that were used by the district to improve the district's elementary reading plan implementation.

First, the rich qualitative data allowed us to determine that there were no differences in the implementation of Open Court or the quality of pedagogy provided by teachers to students in both treatment and comparison classrooms. This allowed us to confirm that our two groups—treatment and

comparison—were comparable in terms of the primary instruction that they were receiving. This, combined with the fact that student achievement was also the same, indicated that the Waterford program had no added benefit for students in treatment classrooms. Furthermore, the qualitative data allowed us to determine that the quality of pedagogy and Open Court implementation in both treatment and comparison classrooms was low. We were able to demonstrate the specific ways in which teachers did not provide sufficient primary reading/language arts instruction to their students by including in our report verbatim text of classroom observations. We then made recommendations regarding how the district might improve the implementation of its primary reading program, Open Court, in addition to any recommendations we made that were directly related to the implementation of the Waterford program. The district responded to this finding by increasing its focus on improving Open Court implementation through a variety of professional development mechanisms.

Second, the rich qualitative narratives allowed us to determine that student engagement fluctuated both during teacher-directed instruction and during Waterford program time. Again, this led us to conclude that the students' instructional experiences were comparable in the two groups. Students were as likely to be disengaged when using the Waterford program as they were during Open Court time, thus demonstrating that the use of technology for instruction did not guarantee that students would be engaged. This finding allowed us to make a number of recommendations to improve the implementation of the Waterford program. We suggested that the computers be positioned within the classroom in such a way as to reduce the likelihood that they would distract students during their Open Court instructional time and to increase the likelihood that students using the computers would not be distracted by each other or their classmates. We also recommended that the computers be moved to a lab so that all students could work on the Waterford program at the same time with teacher supervision.

Third, we were able to use the detailed narrative to determine that the Waterford program was supplanting the Open Court instructional time instead of supplementing reading/language arts time. During the first year, we noticed that there was an instructional time overlap between the time students spent using the Waterford program and Open Court Green Section instruction. As mentioned earlier, we made changes to the computer log protocol to specifically examine the extent of this overlap. The data gathered during the second year allowed us to determine that the length of the reading/language arts block and the policy governing the use of Open Court and the Waterford program

put teachers in the position of having to either 1) use the *supplementary* reading program during the *primary* reading program instruction; 2)

use the Waterford program during other portions of the instructional day (e.g., during math instruction); 3) or not have all of their students use the Waterford program on every day. These time constraints also do not take into account other events that interfere with typical daily instruction like [professional development time], assemblies, parent teacher conference weeks, testing schedules, and a host of other events that further limit[ed] the amount of time available during the regular instructional day. Consequently, it was not surprising to find that approximately half of the kindergarten classrooms and two-thirds of the first grade classrooms were observed using the [Waterford program] during Green Section instruction. In other words, on any given day, between 20–31% of students missed all or part of their primary phonics instruction and instead were exposed to the Waterford courseware, the supplementary reading program. Thus, the program was not used to supplement the primary reading program. (Hansen et al., 2004, p. 64)

Our recommendations included limiting the use of the Waterford program to those students who demonstrated the need for additional support; using the Waterford program only with students who needed additional support during the existing intervention programs that took place after school, on weekends, and during the summers; and extending the school day so that all students would be able to use the program on a daily basis for the recommended amount of time. The district responded to these recommendations by providing schools and teachers with new guidelines regarding the use of the Waterford program during instructional time. In addition, the findings provided support for the district's decisions to move to an all-day kindergarten schedule and to not expand the implementation of the Waterford program beyond its initial schools.

## CONCLUSIONS

In 2001, the federal government stepped into the debate over what is to be considered quality research. In the No Child Left Behind Act, Congress set forth the expectation that for research to be considered of a high quality, it must be scientifically based. The use of this term, *scientifically based research*, has created quite a stir in the research and evaluation communities. It has led to an ongoing public discussion of what is meant by *scientifically based research*. Much of the debate has swirled around whether scientifically based research demands the use of experimental or quasi-experimental design. A second and equally important discussion is under way concerning the types of research methods that are to be considered legitimate, appropriate, and in line with scientifically based research designs. We argue, as others have,

that in addition to the use of quantitative methods, scientifically based research designs should employ qualitative methods when the goal of the research is to evaluate program effectiveness. We believe that qualitative methods have the power to explain why and how program outcomes were or were not attained. Thus, the purpose of this article is to document the challenges and benefits associated with a large-scale evaluation that incorporated qualitative methods. We chronicled the challenges we faced and the decisions we made in light of those challenges. We also demonstrated that, in spite of the difficulties associated with using qualitative methods on a large scale, there are significant benefits as well. More specifically, by using qualitative methods, we were able to improve the evaluation and generate findings that were meaningful and useful to stakeholders. It is our hope that our experiences can be used by others to inform their work as they seek to incorporate qualitative methods into their large-scale evaluations.

### *References*

- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Technical Report 513). National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.
- Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher*, 32(9), 3–14.
- Chatterji, M. (2004). Evidence on “what works”: An argument for extended-term mixed-method (ETMM) evaluation designs. *Educational Researcher*, 33(9), 3–13.
- Datnow, A., & Yonezawa, S. (2004). Observing school restructuring in multilingual, multicultural classrooms: Balancing ethnographic and evaluative approaches. In H. C. Waxman, R. G. Tharp, & R. S. Hilberg (Eds.), *Observational research in US classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 174–204). Cambridge, England: Cambridge University Press.
- Duffy, G., Roehler, L., & Herrmann, B. (1988). Modeling mental processes helps poor readers become strategic readers. *Reading Teacher*, 40, 514–521.
- Ehri, L. C. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 107–143). Hillsdale, NJ: Erlbaum.
- Erickson, F. (1988). Ethnographic description. In J. Von Ulrich Ammon, N. Dittmar, & K. Mattheier (Eds.), *Sociolinguistics* (pp. 1081–1095). Berlin, Germany: Walter de Gruyter.
- Erickson, F., & Gutierrez, K. (2002). Culture, rigor, and science in educational research. *Educational Researcher*, 31(8), 21–24.
- Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on “scientifically based” education research. *Educational Researcher*, 32(7), 31–39.
- Feuer, M. J., Towne, L., & Shavelson, R. (2002). Scientific culture and educational research. *Educational Researcher*, 31(8), 4–14.
- Greene, J. F. (1998). Qualitative, interpretive evaluation. In A. J. Reynolds & H. J. Walberg (Eds.), *Evaluation research for educational productivity* (pp. 135–154). Greenwich, CT: JAI Press.
- Hansen, E. E., Llosa, L., & Slayton, J. (2004). *Evaluation of the Waterford Early Reading Program as a supplementary program in the Los Angeles Unified School District 2002–2003* (Planning,

- Assessment and Research Division Publication No. 170). Program Evaluation and Research Branch, Los Angeles Unified School District, Los Angeles, CA.
- Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in Journal item: education. *Educational Researcher*, 33(2), 3–11.
- Noga, J. E. (2002, April). *Learning in small classes: Using qualitative methods for evaluation to understand how the process of learning differs in smaller classes* (Paper presented at the annual meeting of the American Education Research Association (AERA), LA: New Orleans.
- Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Pearson, P. D. (1984). Direct explicit teaching of reading comprehension. In G. Duffy, L. Roehler, & J. Mason (Eds.), *Comprehension instruction: Perspectives and suggestions* (pp. 222–233). New York: Longman.
- Piburn, M., & Sawada, D. (2001). Reformed teaching observation protocol (RTOP): Reference Manual (ACEPT Technical Report IN00–3). Arizona Collaborative for Excellence in the Preparation of Teachers, Tempe, AZ.
- Pressley, M., El-Dinary, P., Gaskins, I., Schuder, T., Bergman, J., Almasi, L., et al., (1992). Beyond direct explanation: Transactional instruction of reading comprehension strategies. *Elementary School Journal*, 92, 511–554.
- Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction*, 17, 1–24.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15–21.
- Slayton, J., & Llosa, L. (2002). *Evaluation of the Waterford Early Reading Program 2001–2002: Implementation and student achievement* (Planning, Assessment and Research Division Publication No. 144) Program Evaluation and Research Branch, Los Angeles Unified School District, Los Angeles, CA.
- Woodcock, R. W. (1998). *Woodcock reading mastery tests: Revised normative update, examiner's manual*. Circle Pines, MN: American Guidance Service.
- Yopp, H. (1992). Developing phonemic awareness in young children. *The Reading Teacher*, 45, 696–703.

JULIE SLAYTON is a chief educational research scientist for the Los Angeles Unified School District. Her research focuses on the relationship between district-provided professional development for teachers and changes in the quality of teacher pedagogy. Her earlier work focused on charter schools. She is the author of “School Funding in the Context of California Charter School Reform: A First Look” in *The Multiple Meaning of Charter School Reform* (Ed. A. S. Wells), Teachers College Press, 2003, and coauthor of “Defining Democracy in the Neoliberal Age: Charter School Reform and Educational Consumption in *American Educational Research Journal*, 2002.

LORENA LLOSA is an assistant professor in the Department of Teaching and Learning at New York University’s Steinhardt School of Education. Her research focuses on validity issues related to the assessment of English learners’ language proficiency and content knowledge. She is the author of “Assessing English Learners’ Language Proficiency: A Qualitative Investigation of Teachers’ Interpretations of the California ELD Standards” in *The CATESOL Journal*, 2005.