C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Dilemmas and Issues in Implementing Classroom-Based Assessments for Literacy

A Case Study of the Effects of Alternative Assessment in Instruction, Student Learning and Accountability Practices

CSE Technical Report 365

Elfrieda H. Hiebert and Kathryn Davinroy
CRESST/University of Colorado at Boulder

October 1993

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

PREFACE

The current intense interest in alternative forms of assessment is based on a number of assumptions that are as yet untested. In particular, the claim that authentic assessments will improve instruction and student learning is supported only by negative evidence from research on the effects of traditional multiple-choice tests. Because it has been shown that student learning is reduced by teaching to tests of low level skills, it is theorized that teaching to more curricularly defensible tests will improve student learning (Frederiksen & Collins, 1989; Resnick & Resnick, 1992). In our current research for the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) we are examining the actual effects of introducing new forms of assessment at the classroom level.

Derived from theoretical arguments about the anticipated effects of authentic assessments and from the framework of past empirical studies that examined the effects of standardized tests (Shepard, 1991), our study examines a number of interrelated research questions:

1. **What logistical constraints must be respected in developing alternative assessments for classroom purposes? What are the features of assessments that can feasibly be integrated with instruction?**

2. **What changes occur in teachers' knowledge and beliefs about assessment as a result of the project? What changes occur in classroom assessment practices? Are these changes different in writing, reading, and mathematics, or by type of school?**

3. **What changes occur in teachers' knowledge and beliefs about instruction as a result of the project? What changes occur in instructional practices? Are these changes different in writing, reading, and mathematics, or by type of school?**

4. **What is the effect of new assessments on student learning? What picture of student learning is suggested by improvements as measured by the new assessments? Are gains in student achievement corroborated by external measures?**

5. **What is the impact of new assessments on parents' understandings of the curriculum and their children's progress? Are new forms of assessment credible to parents and other "accountability audiences" such as school boards and accountability committees?**

This is one of four reports that document our progress in understanding these questions, based on case studies in three elementary schools.

DILEMMAS AND ISSUES IN IMPLEMENTING CLASSROOM-BASED ASSESSMENTS FOR LITERACY[1]

Elfrieda H. Hiebert and Kathryn Davinroy
CRESST/University of Colorado at Boulder

Assessments that are intended for instructional uses by teachers have extended histories in the field of literacy. The informal reading inventory has been refined by succeeding generations in the form of the running record (Clay, 1985) and miscue analysis (Goodman, 1968), but the underlying concept that teachers sample students' oral reading and retellings and study the strategies apparent in these responses has not changed substantially since Gray introduced the notion in 1920. As Johnston (1984) observed in the first *Handbook of Research on Reading*, informal reading inventories were first used at about the same point in time as the first generation of standardized silent reading tests. Educators and policymakers could have taken one of two routes 70 years ago. The choice of standardized tests and the relative obscurity of informal reading inventories fit the prevailing perspectives of literacy and accountability.

This emphasis (and many would say, over-emphasis) on standardized tests has repeatedly been called into question in recent years as perspectives on teachers, literacy, and assessment have changed. Calls for new assessments that integrate to a greater extent the involvement of teachers and information from classrooms have been heard in both the literacy (Goodman, Goodman, & Hood, 1989) and evaluation (Linn, Baker, & Dunbar, 1991) communities. To date, efforts to create alternatives have resulted in new forms of tests that use long texts from sources that children typically read (e.g., chapters from books, magazine articles), open-ended responses, and questions that encourage thinking (see, e.g., Kapinus, Collier, & Kruglanski, in press; Weiss, in press). While these efforts to date have not integrated classroom-based assessments, they recognize the need for multiple indicators, where information that originates with teachers is used to tell part of the overall picture (Resnick & Resnick, 1989).

---

[1] This paper was presented at the annual meeting of the American Educational Research Association, Atlanta, GA, April, 1993.

While literacy specialists have been involved in the design of these new assessments, many advocates of classroom-based assessments in literacy concentrate on the instructional applications, not the accountability uses (Cambourne & Turbill, 1989; Goodman et al., 1989; Johnston, 1989). Two perspectives can be seen in this literature (Valencia, Hiebert, & Afflerbach, in press). The first places the identification and use of assessments solely with teachers (e.g., Hansen, in press). According to this view, when teachers are given the responsibility for assessment, they will assess what is appropriate, when it is appropriate.

A similar emphasis on instructional uses of classroom-based assessments underlies the second type of project, but a common core of assessments is encouraged in a school or district. Teachers adapt these measures for their contexts, but the assumption is that common goals of an educational community, like fluency or response to literature, can be captured by common, classroom-based assessments like running records or literature logs (see, e.g., Au, in press; Gipps, 1993; Valencia & Place, in press). While the aim of these projects is to inform instruction, a common core of assessments also means that at least some of the information gained from the assessment can go beyond the individual teacher's classroom. Progress on a school's critical goals can be followed by teachers as children move across the grades. Since classroom-based assessments can capture children's efforts on goals and tasks that may be difficult to capture on large-scale assessments, classroom-based assessments have the potential to give educational agencies a broader perspective on student progress toward critical literacy goals.

Studies on teachers' use of assessment have been primarily of the first scenario of classroom-based assessment—where teachers are free to identify assessments. These studies suggest that the typical mechanisms for teacher change result in infrequent use of assessments by teachers. The teachers who were surveyed by Harris and Lalik (1987) used informal reading inventories infrequently, even though considerable time had been devoted to their use in university coursework. Hiebert, Hutchison, and Raines (1991) found discrepancies across two case study teachers in a school where portfolio assessment had been identified as a focus. One teacher used informal reading inventories, studied students' writing samples, and took notes from children's conferences; the other teacher had little information to share with parents,

children, or the research team. She claimed to store information in the "mental filing cabinet." In an extensive examination of teachers' implementations of assessments, Aschbacher (1993) found that teachers were more interested in the activities that could be represented in a portfolio than in the goals represented by these activities. Aschbacher concluded that this focus on activities rather than underlying goals created a barrier to sound classroom-based assessments.

Aschbacher's study (1993), like those of Hiebert et al. (1991) and Harris and Lalik (1987), summarized the manifestations of classroom-based assessment *after* coursework or staff development on classroom-based assessment. The present project was an examination of the issues that arise over the course of staff development on classroom-based assessments. Teachers grapple with many issues whenever they confront their practices, especially when those practices are interwoven with testing and accountability (Anders & Richardson, 1992). An understanding of progressions and issues should be informative to those in schools, districts, and other educational agencies, like Chapter 1, who are contemplating initiating such projects.

Any set of issues or progressions reflects a particular set of structures and contexts. Two features of the particular contexts in this project should be noted. The first is the role of standardized tests. Teachers' concern and attention to their own assessments would be expected to be impacted by the stakes placed on standardized tests. During the year of this project, a moratorium had been called on standardized testing for the third grade in these schools. Consequently, standardized tests did not loom as an issue with these teachers. There was evidence, however, that the participating schools did not regard standardized tests as high stakes. This perspective had some validity since there was no evidence of high demands for test scores from either district administrators or from the state department of education, which has neither curriculum mandates or a high-stakes testing program.

Second, the progressions and dilemmas need to be viewed as a function of the particular workshop content. The project involved a collaboration between teams of teachers at school sites and a group of researchers. Teacher decision making was part of this effort. The intent of the researchers, however, was not to simply verify teachers' existing practices or to discuss possible uses of assessments as has been the manner in the first type of project described by Valencia et al. (in press). The university-based members of this collaboration

began the project with the second perspective on classroom-based assessments. That is, the intent was to work with teachers within a school on a common set of assessments that captured key goals of the school and district. It was expected that teachers would adapt assessments to their classroom contexts. Further, since there were no school, district, or state mandates, teachers had the option of implementing these assessments. The underlying perspective of the research team, however, was that a school would identify some common assessments that captured shared goals. The progression of the workshops was influenced by this perspective in that initial sessions were devoted to the identification of shared goals and appropriate common assessments to capture those goals, and subsequent sessions involved developing scoring rubrics among participating teachers.

Method

Participants

An invitation to collaborate in a classroom-based assessment project was extended to third-grade teams of teachers in a school district that serves the northern region of a metropolitan area. Third grade was chosen because that is the grade at which state and district assessments most frequently occur (Linn, Graue, & Sanders, 1990). A requirement was that all third-grade teachers within a school would agree to participate. Three schools within the same district responded to the invitation and were chosen to participate. Because analyses by socioeconomic class would make it possible to identify individual schools (and jeopardize promises of confidentiality), such an analysis will not be conducted in this paper. However, the three schools represented a range of socioeconomic status, with 62% of the students eligible for the free or reduced lunch in one school and 3% and 9% eligible in the other two schools.

Data Sources

The design of the overall project extends beyond the workshop sessions that are the focus of the current analysis. Accomplishments on a performance assessment are being documented, with baseline and control group comparisons, for all students in classrooms of participating teachers and on teachers' assessments for a representative group of students. Interviews with teachers at three points in the year and transcripts of parent-teacher conferences will be available as well.

Data from transcripts of meetings over September to December of a school year provide the focus of this analysis. While meetings with teachers have extended over the spring semester, these have emphasized assessments with reading of informational text. Meetings over the fall semester were aimed at classroom-based assessments with the reading of narrative text. For two of the schools, seven meetings on literacy assessment occurred during this time period and, in the third school—Walnut (a pseudonym, which is the case for names of all schools and teachers)—six meetings.

Each session was audiotaped and transcripts were made of these recordings. The fieldnotes of a research assistant were used to augment transcripts on features like the material that served as a reference to a teacher's comments and nonverbal cues. Two additional sources—the transcript of a May meeting prior to the fall implementation that was intended for teachers' sharing of their existing assessments, and applications that each school team submitted to the research team—were used to confirm patterns in teachers' views of literacy and assessment.

Foci of the Workshops

The sessions were designed to implement a long-standing perspective on curriculum-instruction-assessment (Cronbach, 1960) that has been adapted for classroom-based assessment (Calfee & Hiebert, 1991). Table 1 presents this overall structure.

Table 1

Foci of Meetings with Teachers

Session	Content of meeting
April 1992:	School team of Grade 3 teachers submit application for participation.
May 1992:	All three school teams meet with research team for debriefing of performance assessment and to share their current assessments.
Session 1:	Shared presentation to all three schools on district goals and possible tools to assess those goals. Teachers meet as school teams to discuss goals.
Session 2:	Research team meets with each school individually to identify particular goal(s) and tool(s).
Sessions 3 to 7:	Research team meets with schools individually on a biweekly basis to discuss through the implementation and use of the assessments. Session 6 included information from teacher-parent conferences.

This framework calls for identification of goals as the first step in the assessment process. Once goals have been identified, appropriate tools that represent those goals are identified. After data have been gathered using those tools, attention turns to the scoring and interpretation of results. In the case of classroom-based assessments, interpretation would lead to consideration of instructional practices.

In line with this framework, the first session began with the selection of goals. The research staff had studied the curriculum framework for the district. The framework reflected the whole language and writing process activity that had occurred in the district over the past seven years. Because the schools came from the same district and the district had a clearly written curriculum framework, all three school teams met together during the first half of the first session. The literacy specialist of the research team gave an overview of the primary goals of the district. Suggestions as to assessment tools/instructional activities that capture these goals were also given. Four goals were emphasized, each represented by a tool/instructional activity: meaning-making—summaries; fluency—running records; interest and extended reading—literature logs; self-assessment—annotations. These four tools had been identified by teachers as representing the goals of a literature-based or whole language program in a previous classroom-based assessment project (Valencia & Place, in press). The research team gave each teacher a notebook in which each goal was described and samples of classroom-based assessments for the tool for that goal were given. For example, after a description of the goal of meaning-making, samples of student summaries from texts that were chosen by students or teachers and teachers' scoring rubrics were provided. After this overview, teachers spent the second half of the workshop meeting as school teams. Their task was to identify the goal(s) for which they would design and implement assessments over the fall semester.

At the next session which occurred the following week, the research team met with teams of teachers at their school sites. This session, for which teachers had several hours of release time, was devoted to identifying particular tools that would capture the chosen goal(s). All subsequent sessions were held on a biweekly basis with a team of teachers from a school. The agreement with the teachers was that these sessions would be an hour in length. Most sessions lasted longer than that, typically an hour and a half. These biweekly sessions

were adapted to the concerns and issues of particular school teams within the framework already described (i.e., scoring and interpretation/instructional implications, following the identification of goals and tools).

Results

Steps of Analysis

There were three steps to the analysis. First, a set of categories was established. Transcripts that represented the three schools and three periods of time—first two sessions, middle three, and final two—were read and notes written down about the content of teachers' comments by two members of the research team. The two individuals would evaluate two transcripts and meet to discuss the categories that they had identified. They would return to another set of transcripts, using the common categories. The category scheme was refined until all of the interactions in the transcripts could be accounted for by the category scheme. This category scheme, which appears in Table 2, was used to code the data for all of the transcripts.

Next, all of the transcripts were coded using this category scheme. Themes were coded as conversations, with the number of turns that teachers contributed to a theme identified. The rounds of conversations for particular themes were clustered together to establish the weight given to a particular conversation during a workshop. Typically, a session had a primary theme (e.g., identification of goals), with a variety of subthemes related to the overarching theme (e.g., how did fluency fit in with meaning-making). The primary themes, with their subthemes, and secondary themes were established for each session.

Finally, these themes were examined as to the presence of a dilemma. A dilemma was defined as an obstacle, perceived or real, in teachers' implementation of assessments. An obstacle implies the presence of at least two divergent points of view—one where the issue is seen to be an obstacle and an alternative perspective where it is not seen as an obstacle (or at least a different obstacle). Often, the alternative point of view was held by the researchers. In some cases (and, as the analyses will show, in one school in particular), dilemmas were apparent among teachers within the school.

Table 2

Categories for Themes and Subthemes

I.	**GOALS**
	A. Defining literacy goals
	B. Relationship to district/state goals
	C. Match of goals and tools
II.	**VIEWS OF ASSESSMENTS**
	A. Standardized tests
	B. Performance/alternative assessments
III.	**LOGISTICS**
	A. Setting up (Preparing materials—e.g., student-selected, same across classrooms)
	B. Execution (Frequency, unit)
	C. Scoring
IV.	**INSTRUCTION/ASSESSMENT RELATIONS**
	A. Assessment event embedded or separate from instruction
	B. Instruction to support strategies being assessed
V.	**INTERPRETATION**
	A. Interpreting student processes/performances based on data
	B. Interpreting based on perceptions of students
	C. Sharing interpretations—parents, children, other teachers.

The analysis indicated that a focus on dilemmas provided only a partial picture of the interactions in and across sessions. Many of the sessions were consumed with discussions among teachers and/or the research team about a theme like how to characterize thorough or solid summaries. Such conversations involved questions and issues but the level of concern would by no means indicate a dilemma. The topics of sessions with no dilemmas are as informative as the presence of a major dilemma. Consequently, the decision was made to develop narratives about the sessions of each school team. These narratives describe the sequence of themes and, when they arose, dilemmas. The narratives also include descriptions of the views of goals and classroom-based assessments held by a school team as a whole. While no attempt has been made to analyze these narratives according to their underlying story grammars, the views of goals and assessments held by a team of teachers can be thought of as the setting and initiating event for the ensuing plot.

Descriptions of Dilemmas and Progression of Themes Within the Three Schools

Pine School. As was true in all the participating schools, the workshop sessions at Pine School proceeded along a particular topical path: establishing goals, determining tools for assessing the goals, implementing the assessments, scoring the products, and interpreting the results. But, of the three schools, the teachers at Pine moved the most rapidly from topic to topic. During a typical workshop, teachers quickly picked up project leader suggestions, worked through tasks, and generally came to some kind of closure at the end of each workshop. Few issues really grew into dilemmas and rarely resurfaced from one session to another. Pine teachers seemed intent on the techniques that project leaders suggested in sessions.

In the first session, teachers voiced a strong commitment to their district framework. One teacher referred frequently to the large binder of district literacy goals as the group deliberated on their project goal: "So are our choices like from 3.1, 3.3 [designations for three major literacy goals], or are they the little sections between them?" (Karen, 9/15, l.13-15). The conversation about goals moved to the conversation that would dominate the remainder of the session: tools for measuring student progress toward the goal. At the end of the session, Pine school's project goal was articulated. Despite the attention to district goals, the teachers described their goal in terms of an instructional technique: "The nine week goal, then, is the story frame and the ending goal would be the 9 to 12 sentence narration" (Sara, 9/15, l.1089-1091).

The emphasis on techniques characterized what the teachers had done in assessment previously. They had used the Gates-MacGinitie (voluntarily since tests were not given until the end of third grade) for placing children into reading groups. While these previous assessment activities were mentioned during the second workshop, the teachers seemed intent on picking up another set of techniques—the techniques related to the running record and written summaries of text suggested by the literacy expert on the project team. They came to the third workshop with assessments completed on most of their students, ready to begin developing scoring schemes. They became the first of the three schools to describe instructional decisions based on the assessments and to describe instructional activities that would foster the processes captured by the running records and summaries. At the third session, Pine teachers

described the use of the new assessments—running records—to change the reading materials for a group of students: "We thought that some of these kids could move up because we think Heath is easier than the ratings of the Gates-McGinitie. So where we have kids placed maybe in the 1.2, we're going to move them into the 2.1" (Elly, 9/29, l.5-9). This conversation surrounding the regroupings then shifted to instructional activities to support the literacy strategies being assessed. Discussion of instructional activities became the focus of much give and take among the teachers with one asking, "Now, so could you finish telling me, what will you do with Sue? She reads it. She blows it. Tell me the follow-up. OK. She blows it. Well, tell me, because I'm not sure what I have to do yet" (Elly, 9/22, l.593-596). Pine teachers speculated on various ways to connect the information from the assessments to instructional activities and agreed to continue addressing this issue in subsequent meetings, "But this is a beginning point, because if we do it at the end, hopefully because of this instruction, they will have improved" (Elly, 9/29 l. 550-552).

These first three sessions moved at a brisk pace at Pine. The next meeting took a different turn. The entire fourth session was devoted to the teachers' concerns about the project. The lead teacher for Pine school had brought a list of their concerns for both the mathematics and reading components of the assessment project. The primary issue for the teachers, with regard to the literacy component, quickly developed into an impasse, a dilemma: how did the assessments fit with their instructional programs? The nature of this dilemma was summed up by the statement of one teacher: "The classroom comes first and extras come later and lately there have been lots of extras" (Lena, 10/13, p. 3 of field notes for workshop). The usefulness of the running records and summaries that had been described two weeks previously for shifting the level of children's reading materials was not mentioned. The teachers perceived that the assessments were taking time from instruction and from afterschool time that should be devoted to instructional planning.

The activity of the next workshop was designed to respond to this dilemma by focusing on the incorporation of running records into everyday tasks. Pine teachers returned to a focus on the task. The theme of "time" was brought up as a tangent of two conversations by the lead teacher of the team in this subsequent meeting, but was not pursued by any of the other teachers again, and did not resurface in any future meetings to the level of the fourth session. Rather, the

nature of changes in instructional practices was a prominent theme in this session. Pine teachers described lessons about summaries and deliberated on appropriate ways to instruct and assess summaries. Several issues arose in this conversation. Teachers spent some time discussing sharing expectations with students, a point raised in an earlier session: "The fact is that kids before they start need to know the criteria. And we lots of times don't do that. I don't" (Elly, 9/29. l.635-637). One teacher raised the issue of using written summaries to assess reading comprehension by describing her scoring scheme then questioning the "two grades that we need to look at for a summary." This teacher went on to observe that "they can have all the information in the story, so you know they have the comprehension which is the main goal, but I picked—I thought we picked summaries at grade level last year because of the writing quality too" (Lena, 10/27, l.728-738). The issue of relying on writing to assess reading comprehension was one of the few issues to resurface in later sessions.

The separation or combining of reading and writing led to another conversation about how to instruct students in the writing of summaries. Instruction practices for the Pine teachers focused mainly on getting students to identify and record story elements. One teacher warned the group about becoming too formulaic: "I think . . . modeling of how they could improve. Because my summaries right now are 'The characters are,' 'The story takes place,' 'The problem, the solution, what happened in the beginning, what happened in the end.' It's so rote, grocery list. It doesn't flow" (Lena, 10/27, l.1923-1928). As the teachers described their instructional activities they begin to grapple with establishing a consistent scoring rubric.

Discussions about instruction continued to dominate the final two sessions of the semester where teachers attended to instructional activities that fostered summarizing and, to a lesser degree, activities fostered by the running records. Large portions of these sessions were devoted to teachers' articulation of criteria for summaries identified as showing "thorough, solid, some understanding, or little understanding." Pine teachers' apparent commitment to a particular type of summary elicited concern about student engagement from one teacher in an earlier session, "I've found I can do it (summaries) only every 2 or 3 weeks because they don't like it" (Lena, 10/13, p. 5 of field notes for workshop). This concern reappeared briefly in this session and led to further sharing of instructional activities. One such activity involved students in scoring

summaries themselves. As Lena explained, "So that's when I started writing terrible summaries and then they would tell me what was wrong with my summaries. But I think that there was some growth because the first week we started they didn't even know what a summary was and if I would have written a crummy one they couldn't have told (me)" (Lena, 11/17, l. 649-654). In discussing and developing a scoring rubric, teachers explored the issue of standards in scoring: "So the reason you choose a, you know, like on the way you're teaching it, you choose 'thorough' because it had all the elements and it was written the way you knew a summary should be written, not because that was one of the best ones out of the class" (Karen, 11/17, l.478-482). Again the conversation returned to instructional activities that can be facilitated by sharing the scoring rubric with the students.

Beginning with their identification of the story frame as the semester's goal, Pine teachers indicated an emphasis on techniques or the specifics of literacy, instruction, and assessment. This interest in specifics meant that they moved quickly in implementing assessments and in discussing the implications of the information gained from these assessments for instruction, such as grouping. But gathering running records on some students and designating events for gathering students' responses to text and then scoring them required Pine teachers to rethink some aspects of their instruction. While this school had been involved in gathering children's work in portfolios, this team of teachers had been using a strand of assessments that occurred as separate events in their classrooms—the Gates-MacGinitie and a commercial informal reading inventory. These teachers had returned to a textbook series for reading instruction, after an extended period in the school where trade books had been exclusively used. The notion of embedding assessments within instructional events was one with which they grappled. The dilemma that these teachers had with the notion that assessments could be embedded in ongoing instructional events came to a head in the middle session. This session seemed to be profitable for both the teachers and the researchers in clarifying the underlying issues and moving to resolution. At the end of the semester, teachers were engaged actively in a conversation as to how instructional practices could support students' responses to text, especially summarizing. In the final session, the teacher who had continued to raise the dilemma of additional demands created by the assessments was able to articulate the discreteness with which they had viewed the assessments early on

in the semester: "I'm sure when we pulled those [pages for running records], we kind of said, we're going to do this for the project so we're going to, throw something together. I mean that's honestly the way we did it" (Elly, 12/8, l.1234-1238). She also was able to reflect on the development that she had seen in students and to recognize her own learning: "It's been a good learning experience for me because I think this has been the best summary writing I've ever had with my kids" (Elly, 12/8, l.1915-1919).

Spruce School. During the first session, the definition of literacy that Spruce teachers agreed on for their semester emphasis was: "The first part is comprehension and not worry about the level and the variety of text. But just whatever they are reading, whatever group they are in, they are understanding" (Tamara, 9/15/92, l.240-244). Comprehension was the emphasis but teachers expressed some reservations about the forms that these responses can take. One teacher stated, "I think a lot of them right now if you told them to summarize a story they wouldn't have a clue what that meant" (Libby, 9/15/92, l.796-800).

This teacher's statement did not mean a reservation about written responses to text, however. Of the three schools, this school had had the most involvement with portfolios previously, with students conducting parent conferences with their portfolios in the spring. The teachers demonstrated this background and interest in portfolio assessment when they brought letters that they have had their students write about books that they were reading to the second workshop. The research team engaged the teachers in extending these assessments with a shared set of texts in a class or across the classes in the school. Teachers agreed with the idea of some designation of a core set of selections, since they were having difficulty scoring students' letters, which were based on individual books. Spruce teachers identified selections in a core group of trade books that covered a range of difficulty from books that one or more of them intended to use over the semester.

By the third session, the majority of teachers had gathered written summaries or oral summaries and running records for most of their students. In line with teachers' concerns with an over-emphasis on fluency, the research team guided teachers in scoring the running records for semantic acceptability (the proportion of sentences that are meaningful). The session was spent productively, with teachers scoring their running records and reflecting on the oral and written summaries of their students. Teachers identified the need to

select another book between *Little Bear* (a primer-level text) to *The Lemonade Stand* (upper second grade) since some children had read the former well but struggled with the latter. During this working session on scoring and identifying books, one teacher questioned the usefulness of this information for instructional purposes: "I can have different reading groups and stuff like that but they're, you know, it's not just reading, it's just universally, kids that just work on a very low level on all there is" (Tamara, 10/6/92, l.1438-1442). This teacher's comment did not consume a primary conversation in this workshop, however, as teachers worked on establishing students' proficiency levels and discussed the text difficulty of the books that they have had students read and summarize.

The differences in views of text difficulty were the source of a dilemma that dominated the fourth workshop. According to one teacher, the assessments went against the school's philosophy: "Our philosophy is kind of to look at the child and help them get from where they are to where they can get during that year" (Libby, 10/20/92, l.229-232). The assessments had created conflicts for teachers. First of all, the assessments made some students aware that they weren't reading well, as indicated in the statement: "I would say she's probably second grade now. But I hate labeling kids, and I hate making them feel bad" (Libby, 10/20/92, l.232-237). Libby went on to identify another problem with this assessment information—sharing of information with parents: "I would be real nervous about using this 'this is where your child is, this is a first-grade level.' . . . I don't want some parent going home thinking 'well, you've got to read to me every night this long because you are only reading at first-grade level' " (Libby, 10/20/92). Libby was the primary spokesperson for these concerns related to interpretation, but other teachers periodically chimed in with "I agree" (Reba, Penny). Tamara identified problems and possible benefits, stating the dilemma quite clearly: "And you know, sooner or later, some of them need to know where our kids are, and some of them need to know that they are reading at pre-primer at third grade. . . . On the other hand, it can be really bad on their egos" (Tamara, 10/20/92, l.367-375). The session ended with no real resolution.

When the research team met with the teachers again, the teachers who had been the quietest during the previous session reported on an instructional response to the dilemma. The assessments had led Penny and Reba to adapt their instructional practices by initiating whole-class phonics lessons: "We decided that we'd do our vowels. . . . So we did start with short 'a' " (Penny,

11/2/92, l.29-32). Libby raised questions about the appropriateness of this instructional practice for everyone: "Some of my kids are reading fifth-grade books with no problem at all. I mean, they can read just about anything that's put in front of them. Why would I want to do anything with short vowels with them?" (Libby, 11/2/92, l.408-414). Libby also described the instruction that she was providing some students: "All I'm saying is the strategies and the things we're talking about in the word patterns, those are the things that I'm doing with my, with two of my groups, but with my other one, we're just doing mostly character discussion and the plots of the story. And, you know, I'm not doing much skills with them and maybe I need to be doing something with them more advanced" (Libby, 11/2/92, 493-503). After an extended conversation that included comments like Libby's about the need for differentiation of instruction based on accurate information about students' proficiencies, Penny equivocated about instruction of short vowels when she looked at her summary sheets of student performances on the running records and written summaries: "Now these guys, this one, she can read anything I give her. So they don't really need that much, just a little bit" (Penny, 11/2/92, l.639-642).

These statements suggest that the assessments had been the impetus for teachers' reflecting and acting on their instruction. They were still grappling with many issues surrounding text difficulty and student proficiency but gathering information on students' performances in their typical instructional tasks had led to changes in instruction. Libby carried the theme of differentiated instruction into the next workshop where she described the benefits of the changes in instruction for the students who are struggling readers: "But I really feel like when I have my groups, I feel like my, my most beneficial group is my low readers 'cause that's when I feel like the most teaching goes on is when I have them in a group" (Libby, 12/1/92, l. 1966-1970). The instructional implications were still being worked out by individual teachers as Tamara's statement indicates: "And there's not time in a day, it doesn't matter if it's my management or it's not my management to get these kids that are as low as they are, up" (Tamara, 12/1/92, l. 2614-2617). During visits to the classrooms at this time, the research team also discovered that Reba who had been quiet during most workshop sessions had been conducting running records almost daily and had been working hard with a handful of students whom the initial assessments showed to be reading poorly.

As part of a conversation during this session, teachers raised the need for more guidance on scoring of summaries and the final session of the semester was devoted to scoring of summaries. Teachers had an extended discussion about their students' written summaries. Unlike the early sessions where the school's philosophy was described as "just whatever they are reading," the discussion indicates high standards for student work as evident in Libby's description of her quandary in giving a high score to papers that are short:

> **I think it is real solid too. It is hard sometimes when I read these because some of them are short and yet they seem to have all the characters and the problems and the solutions and the elements that, you know, and there's others that, you know, go into real detailed explanations about—so it's hard to know, you know, this has all the elements so this is a four, but then someone that added more detail to it, you know, got just a four, too you know what I mean? (Libby, 12/15/92, l.411-420)**

Spruce teachers began with a global perspective of literacy and of assessment of children's progress. The examination of samples of students' oral reading and written responses created dissonance among the teachers. This information was viewed as potentially harmful because it might stigmatize students with the teachers themselves and children's parents. The teachers worked on the assessments, however, and began to discuss changes in their instruction. Some attempted guidance in word-level strategies for the group of students who the assessments had shown were not proficient readers. There were discrepancies among the teachers about how this instruction should occur. However, teachers were examining their instruction from a vantage point other than a view of comprehension regardless of text difficulty. Issues continued to arise but, at the end of the semester, teachers were examining their instructional practices and standards.

Walnut School. The first session at Walnut began with a spirited discussion about goals, and this spirited discussion or "grand conversations," to use Eeds and Wells' (1989) term, characterized ensuing sessions. The grand conversations or debates among Walnut teachers began with their identification of critical goals for the semester's work. This discussion focused on the relative role of fluency and meaning in proficient reading. After hearing comments that emphasized fluency, Judy cautioned that "they were separating the meaning from being able to read fluently" (Judy, 9/15/92, l85-187). A turn or two later in the conversation, however, Abby repeated the view that she and another teacher,

Janet, had been stating: "I just feel like I need to have the kids read and I need to talk to them about what they've read to me that's fluency . . . yes, we're working towards meaning but I don't think that's a goal" (Abby, 9/15/92, l.205-208). And, Jackie brought these differences into the open by saying "I'm hearing two different things" (Jackie, 9/15/92, l. 257-258).

The discussion got heated in the second session where Judy argued against the use of a commercial reading inventory, saying "I give the IRI and I think this is an interesting score, okay? And how does it apply to the book the child is actually reading?" (Judy, 9/22/92, l. 1649-1651). The teachers discussed and resolved the issue by identifying parts of trade books that one or more of them used in their classrooms.

The topics of the conversations changed over the sessions—from defining literacy to developing a rubric that discourages formulaic responses to instructional strategies that foster student thinking. Two characteristics remained constant across the sessions: the form of the conversation and continual references to the underlying processes that the assessments and instructional activities were intended to further. Teachers continually expressed their perspectives and asked questions about the perspectives of their colleagues. As part of these discussions, teachers recognized problems—with their instruction, the assessments, their students' progress to date. The pattern is consistent with Janet's response during the third session: "I wonder if we spend some more time on what a quality summary, that this will get better. Not so much that his reading understanding gets better, but what he's able to express will get better" (Janet, 10/13/92, l.1036-1038). And they worked on finding solutions. During the fourth session, Abby's concerns with the questions on their baseline assessment illustrated this stance: "We haven't scored them yet either because, especially the written, we weren't at all happy with what we got in and my concern and Janet's is that those questions are not getting . . . I'm not getting the information that I can use to tell parents how their kids are doing" (Abby, 10/13/92, l.40-44).

In subsequent sessions, teachers used the identification and implementation of classroom-based assessments to inform their instruction and their understanding of students' strategies. Teachers expressed disappointment with students' progress to date, as Jackie's comments during workshop 5 indicated: "Well, I'm disappointed because I wanted to take their rough drafts of

what they presented to me as a summary and to be able to say, as a class, let's determine a one, two, three, and four. And I don't even have enough to do that. I don't have one summary that I think is good" (Jackie, 10/27/92, l.223-227). In these discussions, teachers considered the nature of the task. Jackie acknowledged that the summaries in question pertained to an entire book and that this was "an overwhelming task for the majority of the kids" (Jackie, 10/27/92, l.78-79). As often happened at Walnut, another teacher jumped in to provide a different perspective on children and their responses: "Another problem, the fact that what we think is important and what the kids think is important is often totally different? They may latch onto some funny little thing that happened in a story that really isn't that important" (Judy, 10/27/92, l.266-269).

Because of a weather cancellation and a schoolwide project that pushed a session into January, the last session of the semester occurred for teachers during the third week of November. Even at this point, Walnut teachers saw differences in their standards for students as a result of their reflections on assessment as illustrated in Janet's comment: "When you tell a kid 'now tell me what you predict' and expect a nice paragraph is much different than seeing a sentence that says 'I predict . . . He will go home.' We're expecting a lot more" (Janet, 11/17/92, l.639-643).

The airing of different views continued as teachers grappled with how to instruct these strategies. Judy raised a view different than that of the other teachers: "I'm almost finding, and I'm not sure if it's true or not, but summarizing seems to be a maturational skill, and some kids have it and other kids don't" (Judy, 11/17/92, l.923-926). Jackie responded to this comment with the observation that "I don't think the kids have had the practice and experience" (11/17/92, l.953-954). Janet also noted that the summarizing that children were being asked to do differed markedly from the retellings that they have emphasized in the past, observing that "we're almost giving kids mixed messages . . . Here you tell me you want me to write everything I think, and now you tell me you don't" (Janet, 11/17/92, l.1029-1036). Janet had identified a critical distinction in the literature on responding to text—the difference between summarizing and retelling. This observation came about through the extended conversations that these teachers had.

If an analysis were done on the amount of description or attention to portfolios in schools' applications to the project, the teachers at Walnut School would be described as the school with the least background of the three on classroom-based assessment. Walnut School's application had a single, passing reference to portfolio use in their school. Unlike Pine, they did not describe prior, schoolwide staff development or workshops on portfolios or, unlike Spruce, implementation of student-led parent conferences with portfolios. What their application emphasized was a commitment to the district framework and the prior work of the school in establishing shared goals. Further, for the research team, the progress of the teachers at Walnut seemed slow at times. The portfolios of their students had fewer written summaries or running records than those of the teachers at Pine at the end of the semester. However, the understandings of underlying processes that the assessments were uncovering and the descriptions of student growth, even in mid-November, were rich. The workshops provided the context for these grand conversations, and these conversations had been acted upon in assessments and instructional activities in classrooms. While providing incentive for these grand conversations about student progress on particular goals, the source of these grand conversations lay in a schoolwide effort over the past year. Walnut teachers began the semester with a foundation. They spent the semester building a solid structure of instruction and assessment on that foundation of perspectives on literacy.

Discussion

The task of embedding assessments like running records and written summaries elicited quite different responses from teachers. School teams had distinct struggles and issues. At all of the schools, working on classroom-based assessments put demands on teachers. At Walnut School, teachers used the context of the workshops to analyze and debate all aspects of an assessment—the materials for running records and summaries, the definitions of scoring schemes, and ways to teach so that students could gain the underlying proficiencies. In the other two schools, the embedding of the new assessments into teachers' instructional programs was the source of at least some level of conflict. The workshops took different progressions in those two schools. Pine teachers, with their emphasis on specifics, moved quickly in implementing the assessments and in using the information. They used the assessments to

develop specific instructional techniques and specific scoring schemes. But, in many ways, the new assessments were seen as add-ons and were not firmly embedded in an underlying foundation. At Spruce, a global perspective on literacy and its assessment created the most dissonance of any of the schools with the classroom-based assessments. The assessments showed Spruce teachers that some of their students were not able to read beyond a very rudimentary level. They wondered if they should share that information with parents or with the students. They worried that this information would force them to label students. They struggled with how to go about guiding students who weren't reading proficiently.

The underlying views of literacy and assessment that a school team articulated during the first two sessions shed light on the particular struggles and issues that arose during the implementation. The broader issue of why definitions of literacy and assessment differed across school teams can be speculated upon. For example, schools have different histories and leadership that at a given point can determine the individuals who are hired. Despite the presence of local histories and issues, these findings highlight several issues for specialists, administrators, policymakers, and other groups as they consider the implementation of classroom-based assessments.

First, attention needs to be given to what is mandated. The negative effects of standardized tests on curriculum have been recognized (e.g., Hiebert & Corley, 1993). To some, the short passages, low-level questions, and the limited responses have been the culprits, not the emphasis on assessment. Change the assessment to tasks that mirror the critical processes of literacy, it has been argued (Resnick & Resnick, 1992), and change the instruction. It is a rare assessment project that begins with a recognition of the transformation of goals and instructional practices in a state or district (e.g., Wixson, Peters, Weber, & Roeber, 1987). The teachers at Walnut who engaged in extended conversations to make the assessments work for them were the only team that had spent time on a schoolwide effort to define goals. The other teams of teachers began with a focus on the assessment techniques—in one case specific (story frames) and in the other global (student self-assessment like student-led conferences).

The project was designed to be an "assessment project" and, early on, there was an emphasis on assessments. But these assessments were emphasized because of the goals of literacy that they represented. Running records can be

used in ways that capture a variety of processes. At one level, running records can be used to provide information on fluency only. At a broader level, however, running records provide information on meaning-making (through children's self-corrections and the semantic acceptability of their miscues). Written responses to text, too, can be used to establish a variety of strategies and goals. A perspective on literacy that can be described as constructivist or sociocognitive (Langer, 1991) was the basis for the selection of these measures. But the perspective itself was not the focus of separate sessions. Rather, sessions focused on assessment techniques. At Pine, the emphasis on assessments meshed with their technique orientation. It may well be that a focus on assessment techniques is a necessary first stage that at least some teachers need to move through to a richer interpretation of literacy, student accomplishments, and assessments.

For the teachers at Spruce in particular, the underlying perspective on literacy—and the notion that text difficulty enters into the reading process—was discrepant with their global orientation. A three-month period is not sufficient to determine whether the implementation of assessments can change teachers' views of literacy and of literacy instruction, but there were indications that Spruce teachers were examining their views about their students' learning and their instruction.

Efforts that attempt to mandate classroom-based assessments should be implemented with an awareness that the responses of teachers may vary greatly. It should also be kept in mind that the teachers in these projects came from the schools that chose to participate in the project (with some indications of "nudging" from one or two of the principals). The willingness of teachers in schools that could not be nudged can only be speculated upon. Staff development specialists and state and district administrators who are looking at classroom-based assessments should be aware that the process is not necessarily one welcomed by teachers. Research efforts like those of the New Standards project will presumably be able to verify the extensiveness and nature of change between the first wave of volunteers and the follow-up cohorts when the assessments are mandated.

Ultimately, the issue of how children's learning is impacted by changes in practices, whether those emanate from new assessments or instruction, needs to be addressed. Future analyses in this project will permit examination of the

ongoing portfolios, teachers' use of information with parents, and accomplishments on a performance assessment. The impact of participating in this assessment project on teachers' work with future cohorts of students also needs consideration. In particular, the nature of teachers' implementations when standardized tests are again administered in these schools should be studied.

The present analysis indicates that the collective view of a school on literacy and assessment manifests itself in different issues and dilemmas over the course of a classroom-based assessment project. As many have argued and demonstrated, regardless of the dimension of instructional or assessment practices (see, e.g., Anders & Richardson, 1992), making changes in classrooms is a long and complex process. Initiating a shared set of classroom-based assessments in schools can contribute to the reflection and debate that underlies this process of change.

References

Anders, P., & Richardson, V. (1992). Teacher as game-show host, bookkeeper, or judge? Challenges, contradictions, and consequences of accountability. *Teachers College Record, 94,* 382-396.

Aschbacher, P. (1993). *Issues in innovative assessment for classroom practice: Barriers and facilitators* (CSE Tech. Rep. No. 359). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Au, K.H. (in press). Portfolio assessment: Experiences at the Kamehameha Elementary Education Program. In S.W. Valencia, E.H. Hiebert, & P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities.* Newark, DE: International Reading Association.

Calfee, R. & Hiebert, E.H. (1991). Classroom assessment of reading. In R. Barr, M.L. Kamil, P.B. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 281-309). New York: Longman.

Cambourne, B., & Turbill, J. (1990). Assessment in whole language classrooms: Theory into practice. *The Elementary School Journal, 90,* 337-349.

Clay, M.M. (1985). *The early detection of reading difficulties* (3rd ed.). Portsmouth, NH: Heinemann.

Cronbach, L.J. (1960). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Eeds, M., & Wells, D. (1989). Grand conversations: An exploration of meaning construction in literature study groups. *Research in the Teaching of English, 23,* 4-29.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Gipps, C. (1993, April). *Emerging models of teacher assessment in the classroom.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Goodman, K.S. (1968). The psycholinguistic nature of the reading process. In K.S. Goodman (Ed.), *The psycholinguistic nature of the reading process* (pp. 13-26). Detroit: Wayne State University Press.

Goodman, K., Goodman, Y., & Hood, W. (Eds.) (1989). *The whole language evaluation book.* Portsmouth, NH: Heinemann Educational Books.

Gray, W.S. (1920). The value of informal tests of reading achievement. *Journal of Educational Research,* 103-111.

Hansen, J. (in press). Literacy portfolios: Windows on potential. In S.W. Valencia, E.H. Hiebert, & P. Afflerbach (Eds.), ***Authentic reading assessment: Practices and possibilities*****. Newark, DE: International Reading Association.**

Harris, L.A., & Lalik, R.M. (1987). Teachers' use of informal reading inventories: An example of school constraints. ***The Reading Teacher, 40,*** **624-630.**

Hiebert, E.H., & Corley, R. (February, 1993). ***A comparison of students' reading on standardized and performance assessments in a high-stakes testing context.*** **Submitted for publication.**

Hiebert, E.H., Hutchison, T.A., & Raines, P.A. (1991). Alternative literacy assessments: Teachers' actions and parents' reactions. In S. McCormick & J. Zutell (Eds.), ***Learner factors/teacher factors: Issues in literacy research and instruction*** **(40th Yearbook of the National Reading Conference, pp. 97-105). Chicago, IL: NRC.**

Johnston, P.H. (1984). Assessment in reading. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.), ***Handbook of reading research*** **(Vol. 1, pp. 147-182). White Plains, NY: Longman.**

Johnston, P.H. (1989). Constructive evaluation and the improvement of teaching and learning. ***Teachers College Record, 90,*** **535-549.**

Kapinus, B.A., Collier, G.V., & Kruglanski, H. (in press). Maryland School Performance Assessment Program: A new view of assessment In S.W. Valencia, E.H. Hiebert, & P. Afflerbach (Eds.), ***Authentic reading assessment: Practices and possibilities*****. Newark, DE: International Reading Association.**

Langer, J. A. (1991). Literacy and schooling: A sociocognitive perspective. In E.H. Hiebert (Ed.), ***Literacy for a diverse society: Perspectives, practices, and policies*** **(pp. 9-27). New York: Teachers College Press.**

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. ***Educational Researcher, 20,*** **15-21.**

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990). Comparing state and district test results to national norms: The validity of claims that "Everyone is above average." ***Educational Measurement, 9,*** **5-14.**

Resnick, L.B., & Resnick, D.L. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), ***Future assessments: Changing views of aptitude, achievement, and instruction*** **(pp. 37-75). Boston: Kluwer.**

Shepard, L. A. (1991). Will national tests improve student learning? ***Phi Delta Kappan, 73,*** **232-238.**

Valencia, S.W., Hiebert, E.H., & Afflerbach, P. (in press). Realizing the possibilities of authentic assessment: Current trends and future issues. In S.W. Valencia, E.H. Hiebert, & P. Afflerbach (Eds.), ***Authentic reading assessment: Practices and possibilities.*** **Newark, DE: International Reading Association.**

Valencia, S.W., & Place, N. (in press). Literacy portfolios for teaching, learning, and accountability: The Bellevue Literacy Assessment Project. In S.W. Valencia, E.H. Hiebert, & P. Afflerbach (Eds.), ***Authentic reading assessment: Practices and possibilities.*** **Newark, DE: International Reading Association.**

Weiss, B. (in press). California's new English-language arts assessment. In S.W. Valencia, E.H. Hiebert, & P. Afflerbach (Eds.), ***Authentic reading assessment: Practices and possibilities.*** **Newark, DE: International Reading Association.**

Wixson, K.K., Peters, C.W., Weber, E.M., & Roeber, E.D. (1987). New directions in statewide reading assessment. ***The Reading Teacher, 40,*** **749-754.**