Teachers' Beliefs About Assessment and Instruction in Literacy

CSE Technical Report 421

Carribeth L. Bliem and Kathryn H. Davinroy
CRESST/University of Colorado at Boulder

March 1997

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
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 policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

TEACHERS' BELIEFS ABOUT ASSESSMENT AND INSTRUCTION IN LITERACY[1]

Carribeth L. Bliem and Kathryn H. Davinroy

CRESST/University of Colorado at Boulder

Educational researchers now recognize that teachers' beliefs and knowledge influence their classroom practices (Borko & Putnam, 1996). The set of beliefs and knowledge that teachers have constructed as a result of their classroom experiences, both as student and teacher, acts as a lens through which they view their practices. This lens can serve to facilitate or hinder teachers' efforts as they set about altering their actions in the classroom, depending on whether and the extent to which their existing beliefs overlap with the philosophical underpinnings of proposed changes to their practices.

Although little research has addressed teachers' beliefs about assessment practices, it stands to reason that their evaluative practices are likewise influenced by their conceptions of what constitutes proper classroom assessment. It also follows that when a reform effort attempts to use assessment as a vehicle for improving instructional practices (e.g., Wiggins, 1989), these conceptions will come into play in determining the paths teachers take. The precise way in which teachers implement new forms of assessment and whether the reform succeeds or fails will depend largely on their beliefs and knowledge regarding measurement and its relation to instruction. Thus, if researchers aim to improve education by altering assessment practices, we must understand the belief system underlying teachers' ways of evaluating their students' learning.

Existing research on teachers' beliefs and knowledge about instruction has treated assessment either not at all or superficially. In instances where it is addressed, a formal definition is implicitly assumed, where assessment is analogous to testing and where its primary purpose is summative in nature (e.g., assigning grades, promoting students). Such a conception, while representing

[1] The authors would like to thank Hilda Borko, Margaret Eisenhart, Scott Marion and Vicky Mayfield for their responses to an earlier draft of this paper. We especially appreciate the guidance provided by Lorrie Shepard.

some narrow slice of assessment, hardly considers all the information-gathering that teachers do as they learn about their students.

This limited treatment of assessment extends to research undertaken within the measurement community. Stiggins and Conklin (1992), for example, describe a series of studies in which they document teachers' classroom assessment practices beginning with a survey study of teachers' self-reported use and concluding with an observation study that detailed the context in which teachers assess their students. Throughout the research, Stiggins and associates define purposes of assessment that are meaningful to the researchers: diagnosing, grouping, grading, promoting students, and evaluating instruction defined as "us[ing] classroom assessment as a means of documenting the success or failure of a particular instructional program" (p. 83). While some of these purposes coincide with teachers' daily use of assessment in instructional planning, little attention is given to the purposes that teachers ascribe to assessment or to less formal means of obtaining information about students.

Indeed, few have attempted to understand teachers' notions about what constitutes assessment and what purposes it serves for them. Airasian and Jones (1993) use a decision-making framework to show how teachers use informal techniques in order to size up their students at the beginning of the academic year, to plan instruction, and to assess the success of that instruction. Even this study, however, catalogues teachers' practices without seeking to understand the belief system that underlies their thinking about assessment. Without considering their perspectives and beliefs, teachers' changes in their assessment and instruction are likely to be both superficial and fleeting.

The purpose of the study reported here was to investigate teachers' existing beliefs about assessment and their connection to instruction in literacy. Because this was our purpose in analyzing data drawn from a larger assessment project, we began this study by defining key conceptual categories that guided our thinking: teachers' beliefs, assessment, instruction, and literacy. Our conceptual framework informed our thinking during analysis about data relevant to these categories.

Conceptual Framework

The Importance of Teachers' Beliefs

Early research on teachers' cognitions demonstrated that their thought processes influence their actions in the classroom (Clark & Peterson, 1986). Teachers' thinking, planning, interactive decision making (the very act of instructing and assessing their students), and implicit beliefs are interwoven facets that impact their classroom practices every day. By extension, then, their implicit theories and beliefs about assessment inform their thinking and planning and, consequently, shape their classroom assessment practices.

Understanding teachers' beliefs and theories about their work is necessary, as Clark and Peterson (1986) comment, in order to "make explicit and visible the frames of reference through which individual teachers perceive and process information" (p. 287). Because these "frames of reference" are tacit, teachers may not be aware of the possible conflict between their underlying beliefs and the philosophical underpinnings of proposed changes to their practices. Yet, these frames provide the organization for their existing knowledge. Further, they act as filters when teachers have new experiences (Borko & Putnam, 1996); teachers selectively filter out information, albeit unconsciously, that doesn't make sense given their tacit ideas and theories about instruction. Because these existing beliefs act as influences on teachers' (and researchers') attempts to change practice, they themselves may be important sites for change as well (Cohen & Ball, 1990).

Specific studies of change have shown that teachers will alter new practices handed down to them so that the proposed changes fit more closely with their existing implicit beliefs (Eisenhart, Cuthbert, Shrum, & Harding, 1988). Therefore, we must make explicit these beliefs and theories in order to change beliefs, change practices, and understand why reforms don't work as they are intended once implemented in the classroom. By making beliefs evident, it may be possible to ease the incorporation of new practices into a teacher's repertoire by highlighting existing differences between her beliefs and the philosophy of the new practices. When dissonances are identified, they can be confronted and debated in a thoughtful way so that teachers choose to retain or alter their beliefs. This identification of conflicting beliefs is especially relevant in the context of assessment reform, which comes laden with its own set of beliefs about what constitutes quality assessment and how it can be used.

Philosophy of the Assessment Reform Agenda

Much has been written of late describing what authentic or performance assessments look like (Resnick & Resnick, 1992) and how their use will promote better teaching and learning in the classroom (Stiggins, 1988; Wiggins, 1989). While reformers vary in the purposes they ascribe to assessments, they generally agree on the characteristics of a quality task. Primarily, such a task requires students to perform in authentic ways, whether they are reading for understanding, solving math problems, or explaining their thinking. The philosophy of the reform movement ascribes other characteristics to good assessments, as well: They demand higher order thinking rather than simple recall of information; they evaluate skills in an authentic context; and they can be embedded in the fabric of regular classroom activities. Indeed, a good assessment is the same as a quality instructional task. As a consequence, teachers can locate their evaluations in daily instructional activities without disrupting the classroom.

According to advocates of performance assessments, such tasks, by documenting student thinking, can serve teachers in ways beyond common measurement purposes. While evaluations are typically used to document student growth or determine grades, the extended responses that accompany students' performance tasks can provide information about errors and misconceptions in their thinking. Teachers can then use this knowledge to tailor instruction to meet the needs of their students.

Collateral benefits of the use of performance assessments abound in the resulting enhanced learning environment. Students become aware of standards for academic excellence as teachers articulate and make public their scoring criteria (Frederiksen & Collins, 1989). Tasks and their companion scoring criteria can be effective tools for communicating about students' abilities and teachers' expectations with parents and other teachers. Well-designed tasks provide opportunities for students to consolidate and contextualize their learning (Crooks, 1988); the assessments themselves act as "central experiences in learning" (Wiggins, 1989, p. 705). Again, the role of the assessment changes from strictly an evaluative instrument to a learning tool. Finally, because performance assessments employ teacher judgment as the primary evaluation tool (Wolf, Bixby, Glenn III, & Gardner, 1990), their use can lead to a "reprofessionalization" of teachers and the teaching workforce.

The Reform Movement in Literacy Assessment

Using performance assessments as a way to focus instructional attention on higher order thinking skills then requires us to answer these questions: What do these higher order skills look like in literacy? and how do we assess them?

As with assessment reform in general, the literacy community puts forth a similar position about tasks appropriate for assessment. These tasks should embody actual student performance—children engaged in the processes of reading (Calfee & Hiebert, 1991; Garcia & Pearson, 1991; Valencia, Hiebert, & Afflerbach, 1994). Yet because literacy experts now describe the act of making meaning from text as a multifaceted task, they recommend using multiple indicators to gauge student understanding. For instance, a teacher might listen to a child read aloud one day, ask her to write a response to the text on another, and then talk with her about what she's read on yet another occasion. This permits the teacher to obtain a collection of artifacts about students' meaning-making of text on which to base her instructional decisions.

In keeping with the notion of embedding assessments in instructional activities, authentic literacy tasks utilize intact texts already present in the classroom. The use of chapter- or trade-books, instead of prepared texts composed from grade-level word lists, also keeps the assessment closer to authentic reading. Tasks as well move beyond literal comprehension questions to elicit interpretation, inference, cross-textual connections—authentic reader responses to text. These oral and written responses are intended to tap the actual processes students use in order to understand and make sense of text.

Such assessments help teachers both understand children's ability to recognize words (using a running record, for instance) and document and encourage meaning making (written and oral responses). They allow for immediate evaluations of understanding because judging a response allows teachers to know whether students understood what they just read by whether they can make connections, predict, and apply knowledge in another context.

The current reform agendas in assessment in general and literacy assessment specifically assert a set of beliefs that may differ from the beliefs of practitioners expected to implement new forms of assessment in their classrooms. Given that so little is known about teachers' beliefs about assessment, it is possible that their beliefs will differ from those advocated by reformers. More importantly, these differences if left unstated are likely to influence the ways in which teachers practice new approaches to assessing their students, especially as they relate to their classroom instruction. A careful study of teachers' beliefs about assessment and instruction may highlight where dissonance exists, so that future reform endeavors can appropriately consider teachers' preexisting beliefs about assessment.

Research Questions

Looking at the intersection of our conceptual categories, this study addresses the primary research question: What are teachers' beliefs about assessment and its connection to instruction in literacy? We consider what beliefs surface as a group of teachers participate in a project aimed at helping them implement performance assessments in their third grade classrooms. Additionally, we seek to understand how these beliefs affect teachers' experiences with new forms of student assessment. This leads to a consideration of how teachers' beliefs differ from the philosophy put forth by the reform agenda in assessment. Finally, we ask what implications teachers' beliefs have for staff development endeavors that focus on alternative forms of assessment.

Methods

The Assessment Project

The data for this paper were gathered as part of a larger project aimed at helping teams of third-grade teachers use performance assessments in their classrooms in the subject areas of reading and mathematics. Researchers solicited the participation of a school district located in a suburb of a large western city, in part because it has curriculum standards that parallel the philosophy held by the research team. Literacy instruction is whole language in nature and can be implemented using a combination of basal texts and tradebooks; word attack skills are addressed, especially for below-grade readers, but in a contextualized manner. In order to take part in the study, individual school teams (all third-grade teachers and their principal) obtained the consent of their parent and school-improvement committees.

In all, fourteen teachers at three elementary schools participated in the study. The research team consisted of an assessment expert, Lorrie Shepard, a mathematics educator, Roberta Flexer, a literacy expert, Elfrieda Hiebert, and a specialist in teacher change, Hilda Borko. Two members of the research team met with the group of teachers at each school for 1-1/2 to 2 hours after school every week during the 1992-93 school year. Workshops alternated between reading and mathematics; thus, every other week, the literacy expert and one other researcher met with teachers in their school teams to discuss issues and concerns related to the use of performance assessments in reading.

Literacy Assessments

When project researchers considered how to remain consistent with the notions of performance assessments and their connection to daily instructional events in the classroom, they recommended that teachers use two literacy tools to assess their students: written summaries and running records. Running records, which are useful for diagnosing and remediating below-grade readers, are administered by following along as a child reads aloud from an intact text. The teacher marks miscues, repeated attempts, and other information about the child's fluency; follow-up comprehension questions or oral retells assess students' understandings of what they have just read. Thus, running records can measure both student fluency and meaning-making of text. In fact, the two skills are related in that slow reading speeds necessitated by focusing on word identification preclude meaningful understanding of ideas in text (Anderson, Hiebert, Scott, & Wilkinson, 1985). Written summaries were recommended for on-grade level readers, who presumably are fluent, as literacy experts believe that summaries require students to relate "the gist of the passage," thus demonstrating their understanding of what they've read (Calfee & Hiebert, 1991). Project researchers offered these performance assessments to teachers as embedded tasks whose purpose was to inform the teachers about their students' reading abilities.

Data Collection

Data analyzed for this paper are drawn primarily from the transcripts of the biweekly meetings. Teachers talking about their experiences as they tried out assessment methods and puzzled over the relationship of those tools to their instruction comprise the central data from which we infer propositions about their beliefs. A secondary data source was a set of standardized interviews, administered to all participating teachers three times during the course of the year-long study. We consider the interviews to be secondary data because of the greater potential for teachers to relate socially desirable responses to our questions about their beliefs and practices.

Data Analysis

Qualitative methods of analysis were used. The authors first coded each of the 45 workshop transcripts (15 at each school) according to the conceptual categories of assessment, instruction and definition of literacy skills. We then looked within these categories to reveal domains important to the teachers, and working iteratively, we collected all included terms that teachers mentioned or implied in the workshops (Spradley, 1979). For instance, a primary domain in assessment was "___ is a characteristic of an assessment system, according to the teacher." Within that domain were terms such as formal occasion, objectivity in scoring, and independent work by students. This process of analytic induction, which kept our findings grounded in the data, allowed us then to organize the domains into a taxonomy that highlighted the relationships between domains (LeCompte & Preissle, 1993).

The process of recursively reading the data and searching, for example, for important qualities of an assessment system facilitated our understanding about teachers' ideas regarding what makes a good instructional program. Within the domain of "___ is a principle of instruction," for instance, teachers spoke of practice, repetition, and variety. By continually referring back to the data organized in the Spradley taxonomy, we eventually identified themes about instruction that mirrored the teachers' key ideas about assessment.

Following Erickson (1986), we used these themes to construct separate propositions that describe these teachers' beliefs about assessment and instruction in literacy. We submit that we have "evidentiary warrant" (p. 146) for these claims about teachers' beliefs as a result of our systematic search through the data for evidence that supported or questioned our hypotheses. Indeed, reviewing the data as described above provided us the opportunity to generate as well as test our propositions.

It is important to note that our findings, while describing the course of events over the project year including changing assessment practices, do not evaluate possible changes in teachers' beliefs. This was a purposeful decision made by the authors for several reasons. First and foremost, the assessment project did not set out to confront or alter teachers' existing beliefs, in part because the researchers began the project expecting that teachers' beliefs were aligned with their own ideas about literacy instruction and assessment. Thus, we do not believe that adequate steps were taken in the assessment project to challenge or change teachers' underlying beliefs about their practices. Also, in keeping with the literature on teacher change, we recognize that sufficient opportunity for reflection and, most critically, time are elements necessary to effect substantial change in teachers' beliefs. The single year of workshops on which this study is based would not be sufficient in time or content to expect fundamental change in beliefs.

Findings

The findings section is organized in two parts according to the two assessment tools. We first relate the teachers' story as they implemented running records followed by a similar account detailing their experiences using summaries. Each narrative, identified with italic type, is a distillation of the teachers' experiences from their perspective, as faithfully as we are able to report. In the narratives, we use quotations from the workshops to bring the reader along on the year-long journey as teachers experimented with alternative assessments. This description is an effort to demonstrate to the reader the complex environment in which teachers operate. Interspersed throughout the narratives are interpretive comments, from the perspective of the authors, in roman type, that illustrate and foreshadow the final propositions that express teachers' beliefs about assessment and instruction in literacy. Pseudonyms are used for all participants except project researchers, who are indicated by their initials.

Running Records—The Story

The story of running records follows a rough chronological sequence of issues arising in assessment development and use. Throughout the year, teachers struggled to analyze and define purposes for assessing student knowledge, and issues were revisited allowing for refinement of earlier perceptions and understandings.

Developing an assessment. *At the first meeting of the school year, teachers at each school agreed that as a primary goal for the first nine weeks they wanted to* *capture "where their students are" in terms of their reading ability. For them, this meant assigning to each child a reading level based loosely on publishers' graded ratings (e.g., 3.1, 1.2, etc.) In response to teachers' broad goals, EH suggested using running records as a way to evaluate both fluency and meaning making. Her initial purpose in proposing the running records was to allow teachers to obtain a "profile of your class" so that teachers could make instructional decisions on the basis of the running record results.*

EH described a "90s version of an IRI" that allowed teachers to use their own reading materials and notation systems, and to embed ongoing assessments within their regular instruction. She recommended that teachers begin using the running record assessment by photocopying passages from books used instructionally and by following along as students read aloud. After teachers became comfortable with the photocopied pages, they moved to using a blank sheet of lined paper, making check marks for words correctly identified. For miscues, substitutions, and repetitions, teachers could make notes on the sheet as students read from intact texts during, for example, a readers' workshop session. Throughout the logistical discussions about how to do running records, EH stressed the informal, embedded nature of the assessment.

During this development period, EH especially advocated using running record assessments on below-grade and struggling students:

EH: You don't have to have all the kids do an oral reading. . . . I'd only have those kids do an oral reading about whom you have some concerns and who you think are below 3rd-grade level. And the other guys, you could have them do exactly the same thing [an oral retell of the story] but in writing. So they would actually do a written summary, "Tell a friend about what happened in this part of the story and what you liked or disliked."

Participants looked for ways to elicit summarized information in authentic ways as in the "tell a friend" prompt.

These themes of focusing on meaning aspects of reading, administering running records only to low-achieving students, embedding the tasks in regular classroom activities, and using assessment outcomes to guide instruction characterized EH's coaching throughout the development of running records as assessments of fluency and meaning-making of text.

One aspect of embedding the running record assessment in ongoing instruction was to begin by using pages copied from complete and intact children's stories rather than prewritten passages designed for reading assessment. In response to questions about selecting appropriate text, EH recommended that each school team of teachers first "level" or rate a set of texts according to the team's knowledge of the books, and from these texts select marker passages for the running record assessment. She asked each teacher to bring five texts that they use in their classrooms to the weekly workshops so the team could negotiate a continuum of increasing text difficulty. Then the teachers could pull, for example, from the list of 2.2 texts to administer a running record. To find where their students were reading initially, teachers planned to administer running records through a sequence of texts representing increasing difficulty until the student failed to perform well.

In order to define "performing well" on a running record, teachers turned their attention to interpretation of completed running records. EH described the analysis as being based on errors in meaning, rather than focusing on word recognition alone. She emphasized meaning as the fundamental criterion for judging students' reading ability:

EH: I'm actually going to suggest that you use a miscue-type strategy. . . . [I]t's how many of the sentences that the kids read are meaningful in terms of the errors that they made.

EH pointed out that running records can supply teachers with several types of information about their students, describing for teachers how to use the records to assess fluency (word correctness) as well as meaning making. EH noted that the oral reading can be used to calculate a "semantic acceptability score" (SAS).

The SAS was calculated by counting up the number of sentences read that make sense and dividing by the total number of sentences in the passage. For example, a child who read, "The children boarded the school bus," when the text read, "The children boarded the bus," would receive credit for that sentence because the inclusion of "school" still makes sense. This is different from a strict fluency score of number of words correct. If one counts up the number of words correctly identified, a percentage score still results, but it solely reflects word recognition rather than meaning. Thus, the running record score, referred to here as the "semantic acceptability score," relayed information about students' ability to understand what they are reading—it addressed more than word identification.

Implementation: Doing it the "right way." *As the development part of the project moved to actual use of performance assessments in the classroom, teachers encountered a number of issues and concerns. An initial concern was how to administer running records "the right way." These queries included questions about the logistics of the assessment itself, such as how many sentences or words to include in an assessment passage, how many questions to ask in the retell, what prompt to use to engage students in their retells, when to ask prediction questions, how to provide a preamble leading up to the target passage, whether to impose a time limit, whether to ask on-grade students to respond to the oral retell prompts in writing. They wanted reassurances about what they considered to be technical aspects of the assessment, as well, including what constituted a passing score, and how strictly to adhere to a 90% cutoff score before they advanced students to the next text level.*

Teachers also worried about specific instances they'd encountered in their first running records, such as how long the student stumbled, how much to coach the student, how many tries to "count" as self-correction. These concerns highlighted teachers' discomfort with respect to running records and calculating the semantic acceptability score.

S: *So you want us to go by sentences.*

EH: *That's right.*

C: *Sentences that make sense . . . if the child stumbles over the word, or they go through it three or four times before they get it, and we're trying to read that, if they finally end up getting it, you wouldn't surely call that fluency of that word, would you?*

EH: *Well, the people who are listening to the kids read are making some judgments as to whether the kids can go on to another level. So there's some qualitative judgment involved there, too.*

Later in the same workshop, another conversation suggests the teachers' confusion about whether they were scoring word recognition or semantic acceptability:

E: *What if you helped them with a word?*

EH: *I think you have to call it.*

E: *Call it what?*

EH: *Well, you have to decide if you could take that word out of the sentence [and have it still make sense].*

The SAS required a judgment call by the teachers and led to questions of "how much sense" a sentence had to make and how to define sense making. Because teachers had difficulty calculating and finding meaning in "semantic acceptability," EH encouraged them to find both possible numerical scores for a running record passage: (a) the semantic acceptability score based on number of sentences that make sense; and (b) a "fluency" score based on number of words correctly identified—fluency, in this case, came to be defined as correctly identifying 90% of the words in the passage.

Teachers' concerns about "doing it right" seem to follow from an understanding of assessment as a formal and important event. Teachers worked hard to formalize the assessment opportunity by deciding on a standardized prompt for the retells, by formulating a standardized protocol (teacher relates preamble, student reads, teacher asks specified series of questions), and by using texts from the agreed-upon list of rated books. They appreciated the numerical accuracy associated with the fluency word-counts and the SAS though they never became comfortable determining or using the SAS. Because the teachers very much wanted to obtain a fair numerical result, they worried when students struggled for prolonged periods of time and repeatedly attempted to self-correct, and they wondered how to reflect those idiosyncrasies in the score.

In October, teachers came to their workshops prepared to relate their experiences with the running records as assessment tools. Despite EH's urging that they did not need to assess on-grade and above students, teachers' first concern was the time necessary to give every student a running record, especially when they found that some students needed more than one running record in order to "top out" or demonstrate their final reading level. They argued that they preferred to spend their limited time administering the read-aloud part of the running record to every student rather than include the retell and assess only some of the students. The consequence of this conviction forced teachers to find ways to shorten the time necessary to administer each running record. As one teacher asked:

T: *When you have the kids do the reading and you're trying to figure out what level they're at, do you have to have them do the little prediction and have to have them tell you the summary? Because that's the time consuming part.*

Other teachers explained that they began to omit the retell at the end of the running record.

E: *See, I don't think we want to use those [oral retells] much anymore.*

LE: *No, I don't either.*

S: *I don't think they're of much value now [because teachers are assigning written summaries to the entire class to gauge comprehension].*

. . .

K: *. . . then you have them tell you what they just read?*

S: *No, I just have them do the reading part of it, not really the comprehension.*

Several facets of teachers' understandings about assessment seem to fold into their decision to drop the oral retell of the running records. Because they believe that fairness demands administration of the same assessment to each and every child, they quickly encountered a problem with time—how could they get around to all 25 students?

Another factor contributing to the termination of the retell was teachers' desire to identify a single, discrete assessment target. When teachers were urged to use the SAS calculation, they interpreted this by narrowing the goal to documentation of word recognition. Thus, from the teachers' perspective, the retell was not integral to a running record score: It was assessing "something else" that teachers considered outside of this assessment activity. Note how S above distinguishes the "reading part" from a student's "comprehension." The written summary task became the vehicle for measuring students' understanding of what they read, a task separate from running records, in both its goal and in its administration.

The end of fall brought parent conferences at each school. At a workshop after these conferences, teachers explained that they shared running records with parents as a way to demonstrate their suppositions about students' abilities as well as to support their instructional choices regarding individual children. For instance,

J: *[The running record] was interesting to share with parents of the kids who were not quite as fluent. To actually talk them through a passage their child had read and to explain to the parent, "This is called a running record and it's just one way for us to record exactly how your child reads the passage. And here are the miscues your child made. We can see that your child is using these clues, or this is happening." I thought that it was very real for parents that they could easily see it. In fact, most of my parents said, "Well, when he reads at home he does the same thing. He doesn't pay attention to the context, or he's*

> *not looking at the ends of words." So that was just like having the kid there reading and then talking about what the child had just read.*

Teachers spoke of the running records' ability to provide concrete evidence to back up their "gut feeling" about students' knowledge and ability. The photocopied pages of text with teacher's check marks and comments furnished proof to questioning parents of their child's reading ability, defined as their ability to recognize words.

Running records' instructional implications. Although a stated goal of the project was to illuminate the connection between assessment and instruction, in general, workshop sessions were not dedicated to discussions about how to connect running record results with instruction.

At the outset of the project and into the fall semester, EH thought that her understandings about literacy were aligned with the teachers' instructional choices about using whole language curriculum. She expected that they would know what to observe and learn from the miscue analyses and how, then, to instruct students to help them master words that they repeatedly missed. However, in light of the running record results, teachers came to workshops requesting guidance about ways they could address the challenges now seen in their students. Because third-grade teachers were unfamiliar with ways to contextualize word attack skills in their literature programs, they were unclear about how to proceed. When workshop discussions then turned to connecting reading instruction with running record outcomes, EH recommended attention to word patterns and word attack strategies—a common aspect of the primary grade reading curriculum.

In January, EH suggested that teachers observe a videotape (Benchmark School, 1989) as a way to think about patterns or families of words. She explained various instructional strategies that teachers could employ in classroom activities like choral reading and using small chalk boards or white boards and markers with small groups of students to explore word patterns. She continued to highlight the connection between fluency and meaning making by asking students to "make sense" of the words and passages they read.

While this response from EH aided teachers in their efforts to improve instruction for below-grade students, it also, perhaps unwittingly, supported teachers' narrower perception that the running record was in the main a measure of word recognition. The instructional strategies that EH provided in response to teacher requests focused on patterns of words, for example finding little words inside the big word that students don't know and examining rhymed word endings. The resulting instruction that teachers described in workshops seemed to further divorce the running record assessment from literacy goals for meaning making.

Refining the assessment. *Despite EH's assurances that they need only use running records on low-reading students, teachers continued to assert and act upon the assertion that, most of the time, they needed to perform every running record assessment on every student. With the accumulation of running record information on every student, teachers explored ways to keep records as a way to display student achievement:*

E: *. . . if we could get it on a master [sheet], (. . .) And maybe so you could do it quarterly. Make sure you get every kid quarterly, and have it, say, if a parent comes in, I mean this is a kind of report sheet on the kid.*

A few teachers spoke so highly of the assessment information they were gathering that they wished they had incorporated a fall running record and a follow-up spring running record on all their students so that they could compare and demonstrate growth in reading ability. This pre/post test model of assessment had its appeal:

JA: *I wished that I had done a running record on all the kids in the fall because I think to start off with that running record and then the summary would have been more valuable and helped me to compare more than just to have this [collection] . . . And if I'd done the running record on all of them then it would have given me that extra piece of information.*

Another teacher explained it this way:

S: *I think it would be helpful to give, not the same test, of course, because hopefully they're past that level, but the same type that we gave at the beginning of the year so you can compare . . . You know, have they grown, not only in their reading ability, but also in their ability to respond to the text? And I think we would see a lot of difference.*

As teachers explored and experimented with the running records, they often referred to a desire for a pretest then posttest model for assessing students' growth.

Workshop conversations around planning and instruction showed that teachers firmly held to the belief that, as an assessment, they needed to perform running records on each of their students every time they used them at all. While they might have been gathering similar information on student performance in informal encounters, as when they conferenced with students, teachers did not discuss these events in terms of assessment opportunities. Such evaluations performed in the course of instructional activities (as sitting with a child and having him or her read for a few minutes) were appropriate as activities tailored to the individual, and thus did not require that "data" be collected on every student. Once the label "assessment" was connected with running records, however, teachers chose to administer them to the entire class of students.

At the midpoint in the school year, EH recommended that teachers expand their selection of texts to include expository materials, explaining that teachers would be "building fluency with different kinds of text." As teachers attempted running records in expository text, new questions arose, primarily about students' familiarity with the vocabulary in expository materials. This dilemma had no easy solution, but the group decided to perform running records on passages containing "neutral" or "common, familiar" topics in order to keep unknown vocabulary from becoming a hurdle in the assessment.

Part of teachers' struggle with running records in expository text again demonstrates their belief that the running record was a measure of decoding skills. The vocabulary in these texts might prohibit otherwise good readers from performing well when they weren't familiar with specialized terminology.

As teachers worked with the running records in expository text, some began to take up EH's message that comprehension and fluency are closely connected though not necessarily markers for each other. One teacher reflected on the beginning of the year:

J: *He picked up* The Boxcar Children *[Warner, 1977] at the very beginning of the year and could basically get through the words, but he had no idea what it said. So we backed up and went into much easier text . . . to get [him] books [that he] can digest. . . . And if you say, "Does that make sense?" enough times, he understands that there is sense-making out of the book.*

Later, another teacher also reflected back:

E: *(talking about a student) Well he's much better about repeating a sentence if it doesn't make sense. I mean, I think we finally have that in his head that you have to read for reading [not for words only].*

As teachers looked back on their students' progress over the year, some noted that word recognition and meaning making had to be connected throughout the instruction and assessment process.

But not all teachers were able to see fluency and comprehension as connected in their assessment and instruction, keeping the focus on one literacy goal OR the other:

> *P: But then when I did a running record on him, with this book, I was just really surprised, you know. He made one mistake. But yet he still didn't understand it too well. I was really confused with him.*

Some teachers, like PM, continued to express surprise that students who could name every word would not understand what they'd just read. Their instructional decisions, which focused students' attention at the word level, reflect this stance and suggest that perhaps some students were learning that reading is simply word recognition and word attack strategies. Indeed, interviews with project students about their experiences with reading assessment (Davinroy, Bliem, & Mayfield, 1994) support this view; while children sometimes indicated that reading is about meaning, they asserted primarily that their teachers were interested in their ability to name words.

Running records propel change in teachers' practices. *As the project year wound down, teachers embodied important changes in their practices that were compatible with the project's focus. For instance, by May, most teachers limited running records' use to their below-grade students, a practice that resulted in their realization that running records provided the most information on below-grade level readers:*

> *L: I think the running record is a real valuable tool especially for the low kids. I've done it with a few of the high kids and it's a waste of time. But with the low kids I think it really does show their growth in reading.*

Yet, even believing that running records made the most sense for a subset of their students, teachers continued to lament not having the time to assess every child's fluency every time.

As another example of changing practices, a few teachers did begin to mention how running records can do more than document student achievement—they can inform instructional decisions:

J: *I would say I know different things in readers as far as the types of miscues that they're making, to be able to instruct for those miscues. For example, I have a student who continually leaves off endings of words and it was very easy for me to notice that after I had looked at two running records that the "Ss" and the "eds" and the "ings" and things like that were being left off. And I don't know that I would have picked it up as quickly on that kind of a miscue.*

However a majority of teachers continued to view the running record assessment rather narrowly, as an occasion for documenting growth, as described by this teacher:

T: *And there is tremendous growth, and . . . his mom's mentioned it to me, and when I read with him, he's using strategies—that's how I can tell if they're getting better. The two kids that I've noticed showing the most growth in my classroom are the ones [whose] strategies stand out when I read with them and do these reading running records.*

Summary of issues relating to running records as assessment. For these teachers, running records' implementation in their classrooms presented some issues derived from their beliefs about assessment. Because teachers held to the general notion that assessment is an official event, they were initially reluctant to perform running records only on selected students. They worried about the time necessary to administer running records complete with oral retells on 25 children. And partly because of this, the oral retell was dropped from the running record protocol.

Teachers also wanted each assessment to have a discrete "target" or goal; this belief cemented the omission of the retell. For them, running records were largely measures of word identification—the retell strayed outside the purpose of assessing students' ability to recognize words. Additionally, suggestions made in workshops for instructional remediation focused on word attack skills and buttressed teachers' narrower view of the running record as an assessment of word recognition only.

Teachers' requirement for objective scoring of students' performance led to quantitative word counts and filtered out efforts to pursue meaning making as part of the assessment. Without tangible evidence of sense making, another criterion for a sound assessment, the teachers couldn't feel confident that they were objectively grading student work. Counting words identified correctly was a more impartial evaluation, and the pages of check marks and miscues documented students' abilities.

Written Summaries as Assessment

The story of using summaries to assess reading comprehension is interwoven with the running record story. As described earlier, the development of running records led to the separation of fluency and meaning-making literacy goals. Because teachers used running records to assess fluency, they sought a different assessment for meaning making. The development of written summaries as an assessment of meaning making from text proceeded along a path parallel to the development of running records. The steps included development of the assessment tool, first attempts to try out the assessment with their classes, followed by instructional responses and refinement of the tool. However, unlike the running record, the use of summaries involved much more extensive development of scoring rubrics and recordkeeping systems. While the process was similar, the different nature of the two assessments led to somewhat different responses and highlights different aspects of teachers' knowledge and beliefs about assessment in general.

Developing the assessment. *In early workshop sessions, when EH advocated using running records to assess both fluency and meaning making and stipulated that running records might be most appropriate for below-grade-level and struggling students, she offered a different assessment for on-grade and above-grade students: the written summary of the read text. EH described her vision of teachers circulating during reading time administering running records while some students read and perhaps others wrote summaries upon their completion of books. EH's emphasis was on incorporating these assessments into the daily events of the classroom. The teachers were concerned, though, about the idea of different assessments for different children.*

J: *So, we will need to do one-on-one [oral retell] with everyone.*

EH: *No. We've actually been saying that above the third-grade level, you guys could actually do the [summarizing] in writing.*

Though EH encouraged teachers to assess different children in different ways and when the occasions arose naturally, the teachers held to a choice of creating a meaning-making assessment for all students that would parallel the running record assessment of fluency. This separation was in keeping with their concern that each assessment tool focus on a single literacy skill or strategy and reiterates their beliefs about evaluating each child every time.

During early workshops, the teachers explored various forms that summarizing could take and considered ways that students could summarize their reading in authentic ways.

L: *Summar[y] is the tool: open-ended, structured formats. There is a whole list . . . of things that we can do to show that.*

T: *Well, you know, like [writing] the letter. . . . [Or] we'll do book jackets where you have to tell on one side what . . . the book was about. Write a friend and tell them if you recommend a book. If they can't do those sorts of [summarizing] things, then they probably didn't comprehend.*

Despite the teachers' early consideration of alternative and authentic ways for students to summarize what they made of their reading, they quickly decided to select one "version" of written summaries. The form this summary finally took was to focus on literal representation of the story with special emphasis on the story elements of character, problem and solution, and setting.

C: *I think lit logs and responses assure showing more on how you enjoy and appreciate rather than "Can I get some meaning?"*

Teachers agreed that when asking students to demonstrate their comprehension of text, summaries would be separate from literature logs and responses and would be strict representations of the content of the text.

Teachers' agreement on one version of a written summary, complete with a list of necessary elements for a quality response, seem to stem from their ideas about discrete assessment targets and issues of fairness. A key component of a sound assessment tool is its standardization in administration and scoring, both of which are simplified by these up-front requirements.

Initial implementation. *As teachers made their first classroom attempts with written summaries, the fairness of using writing to assess reading quickly became an issue:*

T: A couple of my kids, you can't comprehend their . . . writing at all and they can't read so then it is a double whammy. If they have difficulty reading . . . it's going to be really difficult for them to get a letter out to someone about it.

P: Right. Well, then, you'll have to do it orally or have someone else write it for them.

T: Yeah. Just take it down instead.

In another workshop:

E: [H]e understands, he knows what he's supposed to be doing, but his written work is horrible. But if I asked him what the story was about, he could tell me all the details. And then I have others that have very good responses. So then, are you testing written work or are you testing reading ability?

The reading/writing dilemma surfaced often throughout the year of workshops. Part of the dilemma stemmed from teachers unconsciously shifting their focus on summary as assessment of reading comprehension to summaries as an end in themselves. Thus, summaries became tasks that were assessed as examples of "good summaries" rather than as examples of student meaning making of text.

Instructional response to the assessment. *Unlike running records, teachers were better able to move quickly to construct an instructional response to the "demands" of the summary. The teachers agreed that if a summary was to be used as an assessment, it would only be fair to "teach" students how to do well on the test.*

LE: If they don't know how to use these [summary] papers, I can't ever assess them. It's the same thing—teaching kids how to do CTBS because we've found that they were not failing the test because they did not know that knowledge but instead they were not test-wise. I feel the same way with a lot of these strategies [using summary as assessment]. It is not that they are not showing us what they know, they don't know how to do these activities. That's why I have a bunch of samples . . . I'm just a believer that the more modeling the quicker they catch on.

While most teachers focused their instruction primarily on acquainting students with the format of the assessment, a few decided that writing summaries was itself a meaningful literacy goal, worth the instructional time because it integrated reading comprehension with writing skills.

LS: Is [the student] struggling with the format? Do you want to spend time teaching them to do this kind of "test?"

JU: I think we want the kids to be able to do this, but I don't think yet they have the skills.

A: Some of them have never been asked to do it.

JA: These kids don't know how to write summaries . . . I really think that there is maybe something I need to teach and we need to talk more about what makes a good summary.

A majority of the teachers perceived summary writing as a very difficult task for third graders and explored ways to use the summary as a "Fair" assessment of meaning making. The argument went like this: If students don't know how to write summaries well, how will we be able to tell whether they understand what they have read; if students write poor summaries, we can't be sure if they do not comprehend what they have read or if they can't communicate in writing.

L: I think summaries are hard especially for third graders. Look at their writing. They have to write every detail.

T: [Y]ou really have to map it out completely for them . . . And you just have to take it step by step or otherwise it is totally overwhelming.

The idea of "step by step" played out in an instructional continuum that led students from rudimentary story maps, to sentences in a brief paragraph form, to paraphrased emphasis on main events with just enough detail to be interesting to the reader of the summary. The teachers' interest in selected inclusion of detail in the summary relates to the writing skills for good summaries.

L: It is hard sometimes when I read these because some of them are short but yet they seem to have all the characters and the problems and the solutions and the elements . . . There's others that, you know, go into real detailed explanations about [the story]—so it's hard to know.

Concern and interest in the writing skills necessary for a good summary led to instructional approaches that emphasized modeling and whole class discussions about appropriate use of detail and paraphrase. Another teacher noted the predicament of quantity versus quality of the student work:

LE: If they just go through and answer list-wise, it's not going to be a good summary . . . if they just think quantity . . . then that's another way of . . . not doing a good summary.

As teachers encountered these difficulties with their students' summary writing, they set aside instructional time specifically to work on the summaries as part of their reading instruction and writers' workshop time. In project workshops, teachers described adapting classroom management, organization, and assignments to address the goals of summary instruction. They shared strategies for their students where children were scaffolded by the teacher and by one another through activities that included collective writing of summaries, class critiques of teacher-written summaries and professionally-written models, small-group summary writing, and peer reviews of student work.

These efforts at creating instruction in support of the assessment suggest a strong desire on the part of the teachers to "teach to the test." In this case, the test is the writing of summaries, and teachers felt an obligation to help their students be successful. The teachers' general instructional focus on summary writing assured that every student would be prepared to write summaries that demonstrated their understanding of the text. And every student was engaged in writing summaries, regardless of their reading level, because uniform assessment is fair assessment. Throughout the process, teachers asserted that they were measuring comprehension skills while, in fact, their instruction and assessment both included writing quality as a goal.

Refining the assessment: Scoring rubrics. *As teachers developed instructional activities around summaries, they were able to come to workshops with samples of student work that they could use to develop scoring rubrics. EH offered a "method" for developing scoring rubrics that involved teachers in sorting their students' papers into four piles based on a quick read of each paper. Once papers had been arranged into four piles, teachers and EH worked to describe the qualities and characteristics of each pile and to identify anchor papers as exemplars at each level. EH introduced four terms that the teacher could use to describe or rate the quality of their students' summaries:*

EH: I would say [summaries] show Thorough, Solid, Some, or Little understanding of the text . . . a scale of four to one. . . . [What you can] start to do is describe what makes twoness, threeness, or fourness. . . . It would be

> *great if you had a discussion with your kids to see how they define some of this.*

Over the course of several workshops, teachers worked to define the criteria for assessing the summaries. The resulting scoring rubric reflected the aspects of summarizing that teachers deemed important for representing understanding of reading passages. [See Figure 1 for the narrative rubric.]

Development of the scoring rubric permitted the teachers to articulate and establish their criteria for good summaries. Writing a rubric enabled teachers to operationalize and formalize the scoring process and outcome. Indeed, the rubric was tied fairly closely to the story-element definition of summary that had earlier led them to discount literature logs or other literature responses as demonstrations of comprehension.

Creating this scoring rubric raised a discussion of objective scoring and fairness. Many voiced concerns about being sure they all used similar texts, had similar expectations of student work, and could agree on scoring. Without these parallels between their classes, the summaries didn't really seem like assessments to the teachers. For summaries to be assessments, a clear and consistent protocol for administration and evaluation was necessary.

Along with consistency in administration, teachers worried about the consistency of the scoring rubric. They questioned whether rubrics could be generic in some way or whether they needed to construct a different rubric for each text and summary assignment. Their concern was that a rubric would be required for each text so that elements specific to that text could to be stipulated.

4 (Thorough)	Includes the setting, the characters, and plot. Told in an interesting way.
3 (Solid)	May be missing some important parts of the story. May have too many details.
2 (Some)	Doesn't have the most important parts. May have some wrong information. Not told in an interesting way.
1 (Little)	Doesn't make sense. Not enough information is given.

Figure 1. Scoring rubric for summaries of narrative text.

Another issue of fairness and score meaning for teachers was whether rubrics should reflect a criterion standard, rather than a comparative model:

> *J: I explained to [parents] that really what we were doing by putting up that scoring criteria . . . and giving students examples, we felt we were guiding them to what is a good summary . . . and it is true that there are some kids who will never be a four . . . do you agree with that?*

In another workshop:

> *K: Will this scoring rubric maybe change throughout the process? I might get a Thorough tomorrow when I have them do individual summaries that, from that group it's the most Thorough, but then as time goes on somebody might [do better] . . . Or is Thorough going to be set firm the whole time?*

When this issue of normative versus strict criterion standards was raised, teachers agreed that in order to be consistent, they should establish that Thorough is not variable with text, experience, or class profile. This decision seemed to reassure teachers that they would not need to create additional rubrics for summaries and that they could use the rubrics with confidence.

As teachers debated their concerns about objectivity and fairness of scoring, they decided that public sharing of criteria and rubrics was one solution. At all three schools, teachers described encouraging students to refer to the rubric as they wrote summaries. During instruction, they pointed to the rubrics as a guide to whole class summary writing. They also talked about peer and self-assessment using the rubrics as classroom activities.

> *T: [With the rubric] they could hear what made a four, what, you know, they heard from the other kids, too. I'm just kind of intrigued by the fact that they're able to score, they judge themselves so well. They're just right, I mean, it's exactly where I am, really close, most of the time. And that surprised me.*

Using the rubric in the classroom to guide writing and self-assessments as well as peer assessments reassured teachers that they were being fair to students by sharing criteria and giving them practice at the assessments. While this was a new idea for most of the teachers, once given permission to share their criteria, they relished the opportunity to help their students succeed.

Refining the assessment: Using expository text. *In February and March, the shift to expository text placed some new demands on the teachers' beliefs and* *understandings of summaries and of assessment.*[2] *As teachers gathered information about comprehension from students' summaries written from informational texts, they decided they would need a new rubric to assess student performances:*

K: This [the rubric] is going to change . . . I don't think it's set. You're not going to be able to have the characters, the problem and solution. You know, there aren't five parts to this [type of text].

As teachers worked with ways to adapt their existing scoring rubric, they bumped up against questions of how strictly to operationalize the criteria. Several teachers wanted to maintain attention to numbers, as they did with the "five part" summary of narrative, by suggesting, for example, that good summaries of expository text would "contain three main ideas." Others argued that some texts have only one main idea and, if their goal was to assess comprehension, that identifying the "overall main idea" or the "gist" would be adequate.

Teachers were not comfortable making subjective judgments about whether summaries were excellent, good, okay, or inadequate. The urge for a way to operationalize and objectify the rubrics, using numbers, remained a high priority. The new text type, however, meant the "old" way to count success in a summary would not do; a new way had to be found.

Developing the new rubric for assessing summaries of expository text returned teachers to an issue they visited earlier: whether it was possible to create a single rubric they could use for all summarizing assessments, or whether a separate rubric would be necessary for each text type or even for each article students read and summarized.

E: Everything is different. I mean wouldn't you have to have separate criteria to judge if they got the idea or not . . . for each article?

EH urged teachers to look for ways they could create a more generic rubric that would capture what a student's comprehension of a passage meant, rather than worrying about exact details from the articles. [See Figure 2 for the expository rubric.]

[2] For detailed information about how working with expository text affected teachers' ideas of summarizing, see Davinroy and Hiebert (1994).

4 (Thorough)	Organized. Includes main idea and some support. Written in student's own words.
3 (Solid)	Not completely organized, but still flows. Includes main idea and some support. May have copied some phrases directly from the text.
2 (Some)	May not have main idea, but includes some details. OR may have main idea but includes wrong information. May include wrong supporting details.
1 (Little)	Includes a lot of incorrect information. Focus is on unrelated details or only on supporting ideas.

Figure 2. Scoring rubric for summaries of expository text.

As the final months of the project passed, teachers explored ways to use their third-grade animal reports (a district curriculum goal) as sources for knowing that students understood what they were reading. Though they mentioned ways that summarizing can be embedded in instructional activities, they never seemed to develop these activities into assessment opportunities (by constructing scoring rubrics for oral presentations, for example.)

The shift in curricular focus to expository text led teachers to dilemmas between their instructional and assessment beliefs. Whereas narrative texts seem to have a fairly consistent pattern from which to develop a scoring rubric (the character, plot, setting elements), expository texts generate the possibility of needing a specific scoring rubric for each article summarized. In instructional activities, the use of informational text didn't seem disruptive, because summarizing informational text was not used as an end in itself; it was being used as a way to generate information for reports. However, when pressed to use end-products and activities not clearly delineated or initially defined as "summary" or assessment, teachers continued to draw on their beliefs in assessment as different from instructional activities—as fair only if every child performed the same task, and as objective if targeted clearly at a single literacy goal. Thus, the myriad vehicles available for summarizing reading languished as opportunities to evaluate student understanding.

Summary of issues relating to written summary as assessment. Throughout the development and implementation of summary as assessment of reading comprehension, teachers struggled with their beliefs about assessment and how they coincided with the assessment demands placed on them by the project. Though summarizing as a strategy of good readers can be manifest in many ways, teachers needed a consistent format and content on which to assess their students. This became the summary as character, plot, and setting, or, in the case of expository text, main ideas with just enough details to support the main claims. The uniform format of the summary assured that only literal content comprehension was being "measured" and that students' opinions wouldn't cloak a teacher's assessment of student meaning making of the text. Sticking with the literal meaning of the text was one way to assure objectivity as well.

As teachers refined their definition and rubric for summaries, the written product took precedence over the assessment tool. Although for a few teachers summarizing was viewed initially as a valuable writing skill, for most, summaries came to be viewed as an end in themselves and engendered instructional sequences and activities that would ensure student learning and performance. Teachers likened this instructional response to their preparation of students to take standardized tests in the spring. Helping students be test-wise supported their belief that assessments were different from instructional activities and needed special attention because of the stakes involved for their students.

Publicly sharing the criteria used to score summaries also helped fulfill the test-taking instructional motives. Teachers saw such sharing as preparation and practice for the "real thing"—the summary that would be scored and put in the portfolio.

Discussion

Table 1 outlines our propositions about teachers' beliefs regarding assessment and the counterpropositions regarding instruction. As depicted in the table, the two belief systems appear to operate side-by-side without much overlap. Our hypothesis is that these parallel belief systems explain teachers' practices as they tried to meet the demands of the project by using running records and written summaries as alternative embedded assessments. There are, however, interesting situations in which teachers' practices deviated from expected response depending on the particular assessment tool. We now consider each assessment proposition in terms of teachers' experiences with the two assessment tools in order to discuss these cases.

Table 1

Teachers' Contrasting Beliefs About Assessment and Instruction

Assessment	Instruction
1. Assessment is a formal occasional event which has as its fundamental purpose to document student growth in understanding.	Instruction provides an opportunity for practice at acquiring a new skill—generating growth. As such, it is necessarily frequent and ongoing.
2. Assessment provides a tangible product that the teacher can share with student and parents.	Instruction need not produce a concrete product. Teachers often monitor the activity by observation and talking with students.
3. Assessment should evaluate student performance on a single discrete target. A tool can measure fluency or comprehension, for instance, but not both at the same time.	Instructional activities integrate many goals at the same time. Reading group may involve students reading aloud, followed by a discussion of the story.
4. In order for assessments to be fair, they must be uniform, which often leads to standardizing the tasks. Fairness is enhanced by objective scoring and individual efforts.	Instruction is individualized, tailored to the needs of the specific child. It also uses group interactions to facilitate learning for all students.

Proposition 1: Assessment is a formal occasional event which has as its fundamental purpose to document student growth in understanding.

According to these teachers, assessment disrupts the daily routine of their classrooms infrequently because its sole purpose is to provide evidence of students' achievements. This conception of assessment is contrasted with teachers' belief that instruction, which makes up the majority of classroom activity, is an opportunity for students to practice acquiring a new skill. This practice, which is frequent and varied in nature, is necessary as a rehearsal leading up to the assessment of that skill. While assessment provides opportunity to demonstrate learning, the primary purpose of instruction is to allow students to *generate growth*—it is a chance to learn.

Teachers, who initially saw the alternative assessments as rare, big-event happenings, used the information they gleaned from the first running records to group the children for reading instruction. This practice was in keeping with the ideas about the official nature of assessment. Yet, the notion of teaching to the assessment format didn't arise for teachers as they implemented the running records. Our supposition about this lack of influence on instructional decisions is that it reflects teachers' beliefs about what it means to read and how to teach reading. They see the act of a running record, reading aloud, as an authentic skill that one becomes better at just by doing it, and they were initially unclear about how to teach specific strategies evidenced in the running record information. Later in the school year, with help from EH, teachers turned their instructional attention to word patterns as a response to running records. While this effort at remediation closed the loop between assessment and instruction, it also served to reinforce teachers' belief that the running record was a measure of fluency at the word level.

In contrast to running records, summaries-as-formal-occasional-assessments quickly became summary-as-instruction because of teachers' dismay with their initial student products. Many reorganized their instruction in order that summaries could become a primary focus during reading time. This response was in keeping with teachers' beliefs about instruction as an opportunity to practice without risk. Then, when the summary assignment was to be an "assessment," students could count on knowing how to negotiate the format. Because teachers knew how to teach summary from their experience with story maps, they could respond quickly by focusing instruction on that skill.

Proposition 2: Assessment provides a tangible product that the teacher can share with students and parents.

Concrete evidence of performance provides the documentation of student growth. Perhaps teachers possess a heightened sensitivity to this because of the pressure they feel to be accountable to parents and others outside the classroom. They may lack confidence to argue their convictions without the benefit of a physical artifact of student performance.

Instruction, on the other hand, is regularly monitored by observation and casual conversations with students. Teachers talked of their "gut" knowledge of a child's reading ability, raw data provided by their frequent observations and interactions with children during instructional activities. When asked to document that "gut" in the process of listening to children read and talk about what they learned from the text, teachers hesitated. Instructional activities could inform the "gut," but when that same activity was to be an assessment—a test—it needed a more tangible form.

Running records, with their sheets of checks and miscues, provided the necessary documentation that teachers expect of assessment tools. For most of the teachers, this was a new way to keep track of students' oral reading. Yet, oral retells associated with running records disappeared quickly, partly due to the absence of an artifact. In keeping with this belief, teachers wanted a more tangible product, at the very least comprehension questions on sheets that they could fill out when they asked the student about what they had read. (Such a sheet could also serve to standardize the task—see Proposition 4.) Anecdotal jottings, informal notes about what the child said, seemed foreign to the teachers and somehow inadequate as a source of assessment information.

Summaries, in their written form, provided no obstacle with respect to this belief. However, when EH recommended other forms of summarizing (after children complained of having to write still another summary), some teachers acquiesced by expanding their notion of summaries to include oral reports, presentations, and so on. Yet they were reluctant to score such student performances because they offered no paper-and-pencil evidence of achievement.

Proposition 3: Assessment should evaluate student performance on a single discrete target.

Following from this idea, a tool can measure fluency or comprehension but not both at the same time. Yet, teachers design instructional activities in order that students may practice on a variety of instructional goals. For instance, teachers might structure a reading group meeting so that students first take turns reading aloud and then participate in a discussion of the story.

This belief, in part, led to the demise of the oral retell originally associated with the running record. Once teachers began to analyze the running record as a measure of word recognition, they no longer needed the oral retell—it didn't provide any new information about students' ability to name words.

It was also this belief that compelled teachers to use summaries as measures of comprehension. By maintaining a clear definition of an appropriate summary—the focus on literal understanding, including five elements of a narrative text—teachers were certain that students' summaries evaluated meaning making to the exclusion of interpretation, appreciation, or cross-textual goals of literacy.

Proposition 4: In order for assessments to be fair, they must be uniform, which often leads to standardizing the tasks.

The requirements for fairness include standardized tasks administered to individuals (rather than groups of students) and objective scoring devoid of interpretation by the teacher. This is unlike instruction where tasks are tailored to the needs of specific children, are often enacted in group settings to facilitate learning for all students, and require teacher judgment.

Running records were simple to standardize as teachers easily agreed to administer running records from several different levels of text depending on their assumptions about the child's ability. However, as we have seen, the teachers adapted scoring schemes to accommodate their need to be "objective" in evaluating the assessments. Rather than "interpreting" where students' miscues were still semantically acceptable, teachers quickly resorted to counting the number of words correctly identified as the scoring criterion. And, as discussed earlier, teachers dropped the oral retell in part due to their inability to gauge student performance so "subjectively."

Summaries, too, presented problems regarding fairness in the teachers' eyes. In order to arrive at a fair and standardized rubric for scoring summaries, teachers needed to know exactly what they were trying to assess. If a summary needs to have five elements in it, the assignment of a grade is easy by counting up the number of elements that are present. Consequently, the instructional implications were that summaries came to be taught as a rule-bound product. Teachers urged their students to learn the five elements and then include them in a sequential pattern.

Another facet of this belief that was apparent in both assessment tools is demonstrated by the persistence with which teachers held on to the idea of administering a standard task to each and every child every time. From the beginning of the project, researchers encouraged teachers to administer running records on low-ability readers and to allow on-grade students to respond to text in writing. Yet this categorization of assessment tool, depending on student ability, was incompatible with teachers' operating belief system regarding assessment.

While researchers made much progress at convincing teachers to use running records on low readers, teachers remained firm in their wishes that they could (had) administer(ed) running records on the entire class at key points during the school year as a way to demonstrate progress for the individuals. Throughout the year, though, teachers expected written summaries of all students—including below-grade students—not just those exempt from the running record.

The four sets of propositions show how the teachers in this study maintained separate belief systems about assessment and instruction. Notably, while their beliefs about literacy *instruction* mesh with experts' ideas about what it means to learn to read and write, their assessment beliefs continue to be grounded in a traditional view of classroom practices. Because these belief systems coexist without overlap, efforts to embed assessments in instructional activities will be thwarted.

For these teachers, a task is either an assessment or an instructional opportunity—it cannot serve both purposes. Indeed, the purpose ascribed to tasks is at the heart of this matter. Teachers, believing that assessment has as its fundamental purpose to document student performance, cannot conceive of an assessment as belonging in an instructional environment where practice without risk is the goal.

Conclusions

The findings in this study illustrate many tenets of contemporary research about teachers' beliefs. These teachers' existing beliefs about assessment acted as an interpretive lens through which they viewed information about new classroom practices. For example, teachers' belief that separate assessments measure student performance on distinct targets affected their use of running records such that the assessments became evaluations of fluency to the exclusion of meaning. The use of summaries was impacted by this belief as well when teachers held to a literal, standardized format in order to obviate confounding issues of text interpretation or appreciation—instructional goals that they said they evaluated by other means. Not surprisingly, these teachers unconsciously and with the unwitting support of staff developers subverted project efforts by altering the new assessment tools so that their use fit more closely with teachers' personal existing beliefs.

These examples identify some of the existing incongruities between teachers' thinking about assessment and the reform philosophy advocated by project leaders. While teachers held to a traditional and rather narrow view of assessment, project leaders espoused a performance-based assessment philosophy that coupled instructional practices with assessment opportunities. Had these dissonances been identified early in the project perhaps they could have been debated and considered in a more meaningful way. Such a public statement of beliefs and the subsequent debate might have informed teachers' thinking so that they used the new assessments in ways that support reform. By making tacitly-held beliefs explicit, teachers could have chosen to change them or keep them, but they might have recognized the implications such beliefs had for their classroom practice. Without this process of discovery, teachers exhibited only superficial changes in their assessment strategies.

The current reform agenda in educational assessment asserts that a good assessment is the same as a quality instructional activity so that, in literacy, both focus primarily on meaning making. In fact, these teachers portrayed their reading instruction as coinciding with experts' goals for meaning-based learning: They focused on higher order thinking and tried to embed skills in authentic literacy contexts. Additionally, they described using observational and questioning techniques to gauge students' attempts at understanding within an instructional context. Given this compatibility between teachers' instructional belief system and that put forth by reform efforts, it seems like the next step—authentic assessment—should be a relatively easy one for the teachers to enact. However, their ideas about assessment were unconnected from their notions of instruction and so deeply entrenched that this shift was not possible without an intervention that led to a purposeful confrontation of personal beliefs.

Yet another goal of assessment reform involves placing high value on teachers' decision-making abilities with respect to assessments of student ability. Allowing teachers to become responsible assessors calls upon their knowledge of subject matter, pedagogy, and children's learning, as well as how to communicate that knowledge to parents and other teachers. One consequence of teachers' judgments taking center stage is the enhancement of their professional status. If staff development efforts want to empower teachers in this way, then such programs must help teachers become reflective about their practices and beliefs, including the confrontation of potentially conflicting beliefs.

Implications for staff development endeavors that use assessment as a way to improve instruction must begin by taking into consideration teachers' preexisting beliefs about assessment. Because of teachers' personal experiences in the classroom, their ideas are likely to be in conflict with those that provide the basis for reform efforts in both instruction and assessment. As we have seen, if such dissonance is in place, resulting implementation of new tools is liable to look different from and even contradict the vision of reformers. For reform to move ahead, dissonance must be identified and confronted; only in this way can one expect substantial change to occur.

Finally, this study highlights the need for teacher educators to make explicit the connections between instruction and assessment. Too often, it seems that staff development projects are about "assessment" or "instruction," with little attention to how intimately one affects the other. Richardson (1994), in a year-long study of teachers' instructional practices in literacy, found herself inundated with comments and questions about assessment. In a similar fashion, the researchers on this project set out to work with teachers on their assessment practices and were somewhat taken aback by the teachers' requests for assistance in their instruction. If one objective of reform is to link instruction and assessment as they play out in the classroom so that one informs and is informed by the other, then staff development endeavors must explore and encourage that link. Otherwise, it perpetuates teachers' ability to hold and enact separate and distinct belief systems about instruction and assessment.

References

Airasian, P. W., & Jones, A. M. (1993). The teacher as applied measurer: Realities of classroom measurement and assessment. *Applied Measurement in Education, 6*, 241-254.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers*. Champaign, IL: Center for the Study of Reading.

Benchmark School. (1989). *A word identification and vocabulary development program* (Videotape). Media, PA: Benchmark Press.

Borko, H., & Putnam, R. (1996). Learning to teach. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 673-708). New York: Simon & Schuster Macmillan.

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

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 255-296). New York: Macmillan.

Cohen, D. K., & Ball, D. L. (1990). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis 12*, 331-338.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research 58*, 438-481.

Davinroy, K. H., Bliem, C. L., & Mayfield, V. (1994, April). *"How does my teacher know what I know?" Third-graders' perceptions of reading, mathematics, and assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Davinroy, K., & Hiebert, E. H. (1994). An examination of teachers' thinking about assessment of expository text. In *43rd Yearbook of the National Reading Conference* (pp. 60-71). Chicago: National Reading Conference, Inc.

Eisenhart, M. A., Cuthbert, A. M., Shrum, J. L., & Harding, J. R. (1988). Teacher beliefs about their work activities: Policy implications. *Theory Into Practice, 27*, 137-144.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 119-161). New York: Macmillan.

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

Garcia, G., & Pearson, P. D. (1991). The role of assessment in a diverse society. In E. H. Hiebert (Ed.), *Literacy for a diverse society* (pp. 253-278). New York: Teachers College Press.

LeCompte, M. D., & Preissle, J. (1993). *Ethnography and qualitative design in educational research* (2nd ed.). San Diego: Academic Press.

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

Richardson, V. (1994). *A theory of teacher change and the practice of staff development: A case in reading instruction*. New York: Teachers College Press.

Spradley, J. P. (1979). *The ethnographic interview*. Orlando, FL: Harcourt Brace Jovanovich.

Stiggins, R. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan, 69*, 363-368.

Stiggins, R., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany: SUNY Press.

Valencia, S. W., Hiebert, E. H., & Afflerbach, P. (1994). *Authentic reading assessments: Practices and possibilities*. Newark, DE: International Reading Association.

Warner, G. C. (1977). *The boxcar children* (L. Kate Deal, illustr.). Niles, IL: A. Whitman.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

Wolf, D., Bixby, J., Glenn III, J., & Gardner, H. (1990). To use their minds well: investigating new forms of student assessment. *Review of Research in Education, (17)*, 31-74.