

MEASURING CHANGE IN THE PRACTICE OF LITERACY TEACHERS

David Kerbow, University of Chicago

Anthony Bryk, Stanford University

Gay Su Pinnell, The Ohio State University

Emily Rodgers, The Ohio State University

Carrie Hung, The Ohio State University

Patricia L. Scharer, The Ohio State University

Irene Fountas, Lesley University

Emily Dexter, Lesley University

June 2007

Submitted for review to *Reading Research Quarterly*.

The work described here has been supported by a Teacher Quality Grant from the Institute for Educational Sciences, R305M040086, “Can Literacy Professional Development be improved with Web-Based Collaborative Learning tools: A Randomized Field Trial.” We are appreciative of this support from IES. All errors of fact, omission and/or interpretation are solely the responsibility of the authors.

MEASURING CHANGE IN THE PRACTICE OF LITERACY TEACHERS

Abstract

We report on a study to develop a reliable and valid tool to map changes in teachers' literacy practice. The Developing Language and Literacy Teaching (DLLT) rubrics are grounded in the work of Clay (2001), Fountas and Pinnell (2006), and the National Research Council (Snow, Burns, & Griffith, 1998). The DLLT also is anchored in a theory of reflective teacher practice (Schon, 1983) and in a model of expertise development. This model comprises moving from experimentation with new methods through achieving procedural automaticity to the expert orchestration of literacy practices to advance student learning. The study presents empirical evidence that directly examines the construct validity of this framework. A field trial was carried out with 78 teachers who were observed on 275 occasions. Paired observations were conducted on 33 of these lessons to assess inter-rater reliability. The internal consistency reliability across the 8 scales that comprised the DLLT ranged from 0.63 to 0.91. On every scale, adjacent category agreement exceeded 90 percent. Inter-rater reliabilities ranged from 0.78 to 0.91. Items analyses were undertaken to examine and validate the hypothesized developmental pathway from novice toward more expert practice. We also found large differences on rubric scale scores between novice and more expert literacy teachers. The DLLT affords both instrumentation for scientific research and can be use as a practical tool for guiding literacy professional development activities.

MEASURING CHANGE IN THE PRACTICE OF LITERACY TEACHERS

The search for effective practices in literacy education has consumed educators and researchers for decades. Today's call for research-based literacy education has infused even more energy into the quest for "what works." This may seem like a straightforward and linear procedure: (1) conduct research, (2) identify effective approaches, and (3) disseminate practices. This simple logic takes dissemination for granted, however—assuming that once practices are "proven effective," large-scale use of them can easily occur.

In fact, when results are detected for a specific teaching practice, the challenge is only beginning. The real issue is putting the results into practice and then "scaling up" to meet the needs of large numbers of students. When something is believed to work, numerous questions immediately arise: Under what circumstances does it work? For whom does it work? And, perhaps the biggest question of all: What do teachers need to know and be able to do in order for it to work for them?

Time and again, the quality of teaching emerges as a more important factor than any single practice (Darling-Hammond, 1996). Even within "standardized" practices, effective teachers constantly make adjustments that give the approach a maximum chance to work. No effective "scaling up" takes place without addressing the issue of teacher quality. Oddly, however, when focusing on issues of teaching expertise, educators and policy makers tend to treat teachers as either "effective" or "ineffective"—a stance that we never take, for instance, toward Olympic contenders or gifted violinists. In fact, athletes and musicians' abilities are seen as developing over time, as these individuals reach for higher and higher levels of expertise. Their coaches analyze their moves

carefully, assess them regularly, and give them advice that results in a continuous refinement of practice over time. In contrast, we know little about the path of progress that teachers take in pursuit of their expertise.

This study was undertaken to create instruments to measure teacher practice development within several specific instructional activities, each designed to develop young students' reading and writing abilities. The specific objective was to develop a reliable and valid tool to map the instructional changes that occur as teachers take on comprehensive literacy practices in their classrooms. Building on prior small-scale research and the clinical expertise of university faculty with long experience in observation and coaching, we sought to develop a set of rubrics that would conceptually map teachers' development both within specific instructional contexts and as teachers orchestrate students' experiences across multiple instructional activities. Simultaneously, we also hoped to achieve an improved set of clinical tools that could be used by school-based literacy coaches to help inform their own professional development work with teachers.

This instrumentation project is part of a larger endeavor to assess the effectiveness of a comprehensive professional education program and to develop and field test the efficacy of adding Web-based collaborative learning tools to further advance this work. This larger investigation involves a four-year field trial currently being carried out in 18 public elementary schools with significant proportions of African-American, Latino, and low-income students. We seek here to examine changes in teacher practice in response to professional development and coaching, and how these modifications in practice relate to possible changes in student learning in these same classrooms over time.

Given this research and practice context, we proceed first to provide background on coaching and on the use of observational instruments in literacy research and practice. Then, we review the distinct instructional components that combine to form a comprehensive framework for literacy instruction. Next, we detail the methodology for examining the reliability and validity of the rubrics in classrooms. Finally, we present the results of our statistical investigations.

Literacy Coaching as Key Context

In recent years, schools have moved to provide sustained school-based, job-embedded professional development. Experts such as Darling-Hammond and McLaughlin (1996) claim that “effective professional development is . . . sustained, ongoing, and intensive, supported by modeling, coaching and collective problems solving around specific problems of practice” (p. 203). Yet, surprisingly sparse evidence exists to document what kind of long-term teacher development actually pays off in terms of student learning.

With the goals of deepening the implementation of research-based practices and, ultimately, raising achievement, many districts are placing literacy coaches in classrooms. In fact, coaching has become a key component in the professional development repertoire of many school districts across the United States (IRA, 2004). It is broadly assumed that a coach’s “technical feedback” is important in helping teachers refine their craft of teaching (Joyce & Showers, 1983). According to Kohler, Crilley, Shearer, and Good (1997), teachers make few changes in their teaching after a typical workshop or training session alone. Informed coaching, in contrast, can make a difference that means going beyond “lists of good teaching behaviors, performed in

mechanical ways in response to supervisor's observations" (Lyons, Pinnell, & DeFord, 1993, p. 43). Among other things, to have an impact coaches need to lift teachers' theoretical understandings. To do so requires:

- deep preparation in the subject matter so that they can make decisions on the "how, what, and when of teaching";
- a deep theory of learning in the particular subject matter area, along with a knowledge of teaching standards; and
- the ability to observe and analyze teaching, with the goal of helping teachers become more self-analytic about their own progress along a developmental continuum of practice improvement.

To achieve these ends, coaches need good tools to guide their daily practice, advance their own learning over time, and, in so doing, improve the quality of the overall professional enterprise.

The Role of Direct Observation to Inform Coaching

As recently as fifty years ago, researchers in clinical psychology began to move away from reliance on the data from individual self-reports in favor of using direct observation, judged to be a more reliable source of information (Hops, Davis, & Longoria, 1995). Individuals asked to self-report on their behavior, for example, often focused on negative experiences and ignored positive ones. On the other hand, as Hops and colleagues noted, direct observation provided more objective evidence from which to interpret behavior.

In educational research as well, direct observation has been used to study teaching since the 1930s when researchers began to explore teacher-student interactions and other

classroom behaviors (Evertson & Green, 1986). It seems likely, therefore, that in order to gain reliable information about how teachers are developing new practices, researchers and coaches profitably might rely on direct observation rather than on self-reports.

What Should Coaches Observe?

From a sociocultural perspective, learning is defined as changing participation in an activity (Rogoff, 1990). This notion of changing participation is anchored in Vygotsky's conceptualization of cognitive development as progress through a zone of proximal development (Vygotsky, 1978). Rogoff (1997), for example, writes about teaching and learning as a process of guided participation through increasingly complex activity. Similarly, Tharp and Gallimore (1988) refer to this phenomenon as assisted performance. A central feature of this conceptualization—and key for our work—is that changing participation in an activity is evidence of learning.

It seems sensible, then, to think that a coach can and should look for changes in how a teacher teaches as evidence of learning over time. For example, a coach who has been working with teachers on book introductions for young children should look for different behaviors before and after coaching.

Prior Efforts at Assessing Literacy Teaching and Its Improvement

Recognizing the need to have firsthand empirical measures of literacy classroom practice, a number of research groups have developed rubrics for measuring teacher skill in generic classrooms (Junker et al., 2005; Sterbinsky & Ross, 2003) and in classrooms comprised of heterogeneous groups (Stanovich & Jordan, 1998), of special education

students (Englert, 1984), and of English Language Learners (Graves, Gersten, & Haager, 2004; Haager et al., 2003).

The research groups mentioned above base their rubrics on a variety of different theories and research literatures. Stanovich and Jordan's checklist was grounded in social constructivism and findings from research on effective teaching. Designed for use in heterogeneous classrooms, their rubric was organized into four categories—Classroom Management, Time Management, Lesson Presentation, and Adaptive Instruction. An instrument by Haager, Gersten, Baker, and Graves (2003), called the English Language Learners' Classroom Observation Instrument (ELLCO), is grounded in reading research described in the reports from the National Reading Panel (National Institute for Child Health and Development, 2000) and the National Research Council (Snow, Burns, & Griffin, 1998). Their rubric includes categories such as Vocabulary Development, Phonemic Awareness and Decoding, and Explicit Teaching/Art of Teaching.

Similarly, Sterbinsky and Ross developed their Literacy Observation Tool (LOT) in accordance with the NRP and NRC findings, and their categories included Instructional Strategies, Learning Environment, and Materials Used. The Instructional Quality Assessment (IQA) instrument developed by a team of researchers led by Junker (2005), drew on research on "best practices" in teaching and on Snow's (2002) work on the five levels of reading comprehension—recognizing and recalling, comprehending, applying new knowledge from a text, synthesizing, and evaluating texts. This rubric focused on three instructional behaviors that they considered most "proximal" to student learning: quality of classroom discourse; rigor of assignments and assigned texts; and

expectations communicated to students about the quality of their work (Junker & Matsumura, 2006).

Each of the rubrics mentioned above use scales that ask observers to assess effectiveness, consistency, or quality. Though they vary in the language used to describe instruction and the specific behaviors assessed, all include some items that refer to teacher talk and/or student-teacher interactions. Stanovich and Jordon, for example, included an item that assesses whether the teacher “evaluates students’ understanding of seatwork tasks and cognitive processes by asking students ‘what, how, when, why’ questions related to the targeted skill or strategy” (p. 233). Sterbinsky and Ross’s LOT evaluates whether the teacher “asks students for predictions” (p. 13). Haager’s group assesses whether the teacher “checks students’ comprehension of text by asking questions,” and “engages students in meaningful interactions around text” (Gersten et al., 2005, p. 200). The IGA developed by Junker et al. has the greatest emphasis on student-teacher interaction, with many of their items specifying both teacher and student behavior, such as: “The teacher guides students to engage with the underlying meanings or literacy characteristics of a text. Students interpret or analyze a text and use specific examples from the text and/or cite examples from the text to support their ideas or opinions” (Resnick & Junker, 2006, p. 38).

Most of the rubrics described above were designed to serve both research and practice improvement. The IQA group (Junker & Matsumura, 2006), for example, envisions their rubric being used for coaching, teacher study groups, and teacher self-assessment. They maintain that their scales are useful for professional development because, although teacher behaviors were rated on a 1–4 scale from “poor” to

“exemplary,” each point on the scale includes a practice descriptor. For example, under “Active Use of Knowledge: Analyzing and Interpreting Text (Grades 1–2),” level 1 is described as “The assignment task guides students to recall isolated, straightforward facts about a text *or* write on a topic that does not directly reference information from the text,” while level 4 is described as “The assignment task guides student to engage with the underlying meaning or nuances of a text. The assignment task guides students to interpret or analyze a text and use extensive and detailed evidence from the text to support their ideas or opinions” (p. 7). Junker and Matsumura write that “the lesson observation and coding protocol coupled with the precise descriptors offered by IQA rubrics may enable practitioners to work in collaboration with colleagues to figure out not only their strengths and weaknesses, but also to identify ‘next steps’ for instructional improvement. Further, the IQA could be used to track growth over time” (p. 12).

The Developing Language and Literacy Teaching Rubrics

The observational rubric developed in our study—the Developing Language and Literacy Teaching (DLLT) rubric—is grounded in the reading theories of Marie Clay (2001) and Fountas and Pinnell (2006), the effective literacy practices identified by the National Research Council (Snow, Burns, & Griffith, 1998), and a theory of reflective teacher practice based on the work of Schon (1983). Like the IGA, the DLLT specifies behaviors arranged along a continuum of expertise, but it differs from the IGA in several respects. First, though the IGA was based, in part, on research on reading, it was developed as a tool for assessing teachers across subject domains. As a result, the classroom practices described in the rubric are of necessity generic. The DLLT, by contrast, was developed

specifically for assessing literacy and language teaching. It focuses on teaching behaviors within specific instructional activities, such as interactive writing lessons and guided reading lessons, that are core components in most comprehensive literacy instructional systems. Thus, it offers a more micro-level assessment of teachers' literacy instruction. In addition, although Junker et al. envision data from their rubric being used to measure teacher change, the DLLT is specifically based on a theory of teacher development that proceeds from procedural teaching to skillful teaching with expert facilitation of student talk about their reading and writing. Moreover, in the research described below, we present empirical evidence that directly examines the construct validity of this framework for describing the development of expertise in teaching literacy to young children.

Literacy Collaborative as Generative Context

The observational tools described in this paper arose out of our long-standing collaboration around a comprehensive professional development program for literacy educators called the Literacy Collaborative (LC). The Literacy Collaborative was founded in 1993. It brings together universities, school districts, and individual schools in long-term partnerships to improve the teaching of reading, writing, and the other language arts. Specifically, the Literacy Collaborative seeks to integrate a set of instructional practices for reading, writing, phonics, and word study into a comprehensive literacy framework for instruction in elementary classrooms. Learning to develop expertise in these instructional practices, in turn, undergirds the LC professional development initiatives. The work of the Literacy Collaborative is anchored in a deep research base about the core developmental processes involved in learning to read. The

key implementer is the school literacy coordinator. A key component of the literacy coordinator's role is to provide ongoing professional development, including coaching, for classroom teachers.

A perspective on adults learning to become reflective practitioners. The Literacy Collaborative approach to comprehensive literacy instruction is rooted in a model of reflective teacher practice (Schon, 1983). Quality classroom instruction involves an ongoing, interactive process of analyzing students' strengths and needs; selecting specific teaching approaches based on these, continually assessing the results of one's efforts to inform subsequent teaching moves; and maintaining warm and trusting relationships with children while simultaneously engaging their interests (Russell & Munby, 1991; Steiner, n.d.). Teachers need intensive training to carry out such instruction. They also need continuing opportunities to reflect on their teaching and talk about their interactions with students with a more expert practitioner who can support development of the kinds of deep understanding necessary to sustain such a professional practice.

The Literacy Collaborative was designed to deliver this professional development support. At the core of this work is a new professional role—the literacy coordinator. Literacy Collaborative has designed a specific curriculum to prepare literacy coordinators to take on this demanding role. The training includes honing a coordinator's expertise in using the comprehensive literacy framework to teach students and in acquiring new staff developer and researcher skills. The literacy coordinator assumes full responsibility for providing a range of school-based professional development opportunities including group professional development workshops, study groups, and one-on-one coaching.

A perspective on students' literacy learning. The work of the Literacy Collaborative is anchored in a systemic view of students' literacy development.

1. *Reading.* Accomplished readers have acquired a processing system that consists of an integrated set of strategic actions by which readers extract and construct meaning from written language. Readers are engaged in complex thinking that is largely transparent to them as a process. That is, we concentrate on ideas without conscious awareness of what is happening in the brain, which is simultaneously controlling every part of the body. Inside the human brain, processing or problem-solving is taking place. According to Clay, "Processing refers to getting access to and working with several different types of information to arrive at a decision" (2001, p. 80).

We cannot truly separate thinking into compartments. Before, during, and after reading, individuals continually think. Within the text, readers notice the important information and details and putting them together in a coherent way. Beyond the text, readers make their own connections and bringing prior knowledge to reading, and making hypotheses. Readers also notice important aspects about the text such as language, organization, and writer's craft (Fountas & Pinnell, 2006). All these complex operations occur simultaneously as readers meet the demands of texts on the processing system. When we say that a text places demands on a reader, we are really asking: "What must the reader do in order to read this text with understanding?" The answers to that question help us realize what readers are required to do. No matter how simple the text is, the answer is always more than decoding the words.

Readers must draw on a wide range of information to process a text successfully. Some information is *visible*, that is, you can see it in the text. It includes all the symbols

and art in the text. Some information is *invisible*; it exists in the reader's brain. It includes knowledge of the world, content, texts, and language as well as all of the experience the reader brings to reading the text. Texts provide the opportunity and actually demand that readers "mix" visible and invisible information. The reader is constantly processing both kinds of information. The different kinds of information that readers use are complex and wide-reaching. Within a processing system, visible and invisible information work together (Clay, 2001).

2. *Writing.* Young writers, too, are engaged in acquiring an extremely complex process. They are challenged to compose written language using syntax that is different from oral language. Oral language, which comes out in temporal sequence, must be represented with combinations of forms that are arranged in space on the page. Keeping a meaningful message in mind while constructing it letter by letter requires coordination of many different kinds of information—the particular sequence of words arranged in syntax to convey a message; the specific letters needed to form each word (as related in both simple and complex ways to the sounds of language); word parts, sometimes unrelated to sounds, that nevertheless appear as building blocks of words; and the directional movements needed to write the letters in order. Writers are orchestrating a large amount of information.

3. *The reading and writing relationship.* In processing a text, readers actually engage in a wide range of physical, emotional, and cognitive actions. Such processing refers to "all the activities happening in the learner's head, brain, mind or neural networks" (Clay, 2001, p. 124). The processing system utilizes all aspects of relevant knowledge—letter and word recognition, connections to language knowledge, accessing

of content, personal, and text knowledge. All of this involves systems of strategic actions in the brain. When we refer to *thinking*, we mean the in-the-head consideration of the text; as an active process, thinking is just about the same as the term comprehending, but is more inclusive of the wide range of ways readers act on texts and is not so tied to testing and particular curricular approaches and specific texts. Response is an important part of the act of processing and involves many ways of thinking.

Reading and writing are different but highly interrelated and complementary processes. When they are engaged in writing, children are also learning a great deal about how written language works. Conversely, when they engage in reading, they are encountering the syntactic patterns and the forms of written language. When they hear and discuss texts in interactive read aloud, children are gaining vocabulary and learning how texts are structure and organized, which adds to the resources they employ in reading and writing. Teachers who understand these connections and how they can impact student learning can actively and intentionally foster them across instructional contexts throughout the school day.

A General Framework for Comprehensive Literacy Instruction

The core principles discussed above about teacher practice and students' literacy learning constitute the conceptual undergirding for comprehensive literacy instruction. Such instruction typically consists of several interrelated instructional activities, each of which has particular purposes and must be examined as a distinct instructional context. Small-group reading instruction, for example, aims to ensure that students learn to comprehend written texts (Person & Fielding, 1991; Pressley, 1998), as well as learn to use phonics

skills to take words apart while reading for meaning (Pressley, 1998; Snow, Burns, & Griffin, 1998). Through such activity, teachers aim to address comprehension and vocabulary development while simultaneously providing explicit instruction in reading fluency (NICHD, 2001; Pinnell et al., 1995). Teachers also provide daily minilessons on conventions, skills, and the craft of writing. Instruction in writing, in fact, contributes substantially to children's understanding about words (Clay, 1991; NICHD, 2001) as they learn to hear the sounds in words (phonemic awareness) and learn to look at letters and words (Liberman, Shankweiler, & Liberman, 1985; Vellutino & Scanlon, 1987; Lundberg, Frost, & Petersen, 1998) in ways that support both reading and writing achievement.

Effective comprehensive literacy instruction, however, involves more than just a routine enacting of each component. Teachers must establish and maintain a complex social organization within their classrooms necessary for such activities to occur and they must be able to skillfully orchestrate instruction as students move across these component activities. In this latter regard, good teachers strategically “echo” key ideas across instructional contexts to make instruction more powerful within each context.

This perspective directly shaped our efforts in developing observational rubrics for literacy and language teaching. In particular, we chose to organize the developmental rubrics around the six specific instructional components or contexts, commonly found in comprehensive literacy instruction, and around teachers' efforts to organize their classrooms and orchestrate the enactment of these practices in ways that maximize student learning. We describe each of these briefly below.

Comprehensive literacy instruction is typically organized around the following literacy activities:

1. *Interactive read-aloud (usually whole-class instruction)*. Teachers read aloud to students an array of texts that are carefully selected to help students think in various ways about texts. The teacher uses intentional conversation (conversational moves directed toward a goal of instruction) and also promotes routines such as “turn and talk” to help children learn how to talk with each other about texts. The opportunity to engage in “text talk” is rich (Beck & McKeown, 2001). The teacher is decoding the words of the text by reading aloud, but in every other way, young students are processing it and expanding their understanding through talk that is grounded in texts (Fountas & Pinnell, 2006).

2. *Shared reading (whole-class or small-group instruction)*. In shared reading, the teacher and children read from a common text that is either enlarged or on multiple copies. Usually, the text is read several times. Group support helps students to process more difficult texts that they could independently, although it is still important to match the complexity of the text to the group. Using this familiar text, the teacher makes appropriate teaching points that extend children’s understanding of the reading process (McCarrier, Pinnell, & Fountas, 2000).

3. *Guided reading (small group instruction)*. Students are grouped because they are similar in their development of a reading process at a point in time. The teacher selects a text that is appropriate for the group, introduces it in a way that will help students read it effectively, supports individuals during reading as needed, and invites students to discuss it afterward. The teacher makes specific teaching points directed toward any aspect of the reading process (Fountas and Pinnell, 1996).

4. *Interactive writing (whole-class or small-group instruction)*. In this highly supportive context, children can fully participate in the writing process (McCarrier, Pinnell, & Fountas, 2000). The teacher and children collaboratively compose a text and then write it word by word on a large chart. At several carefully selected points, the teacher invites individual children to come up to the chart and make contributions by adding letters or words. These occasions have high instructional value in helping children learn the construction of words (phonics) as well as important aspects of the writing process.

5. *Writing workshop (whole-class and individual)*. The teacher provides a minilesson on any aspect of writing; then students write independently, conferring with the teacher. Then, there is a brief sharing period during which the teacher can reinforce the minilesson principle and invite writer to writer feedback. Students write daily, applying critical principles to their own production of writing in a range of genres.

6. *Word study (whole class)*. The teacher provides a minilesson on phonics and students apply the principle independently. While phonics and word study are embedded in all the previously described contexts, here the instruction is preplanned, direct, and explicit. The emphasis is to teach directly important principles related to how words work as you read and/or rules of standardized spelling (Pinnell & Fountas, 1998).

Orchestrating effective practice. The reading and writing activities described above constitute a repertoire of practices that good teachers weave together based on their pedagogical knowledge and their observation of children. A great deal of this activity occurs simultaneously in a comprehensive literacy classroom and this in turn places considerable demand on the basic organization of the classroom to support effective

instruction. Unless these *general aspects of teaching* are solidly in place, the routines of instruction may easily break down and the intended student learning is then unlikely to occur.

In addition, the effective use of each instructional component entails a complex teaching practice rooted in a deep understanding of developmental reading and writing processes. Rather than being a framework that can be scripted and routinely delivered, teachers must actively design their classroom activities based on evidence about how their students are actually developing as readers and writers. In its most expert rendition, comprehensive literacy instruction requires a thoughtful orchestration of activity across the multiple framework elements described above. Deeper processing, such as comprehending, is not learned in discrete lessons that direct children to practice a “strategy” as if it is one skill to be learned from one book. Rather, teachers continually remind children of aspects of processing across the day and in connection with many texts that they read or write. This *teaching for strategies* (Fountas & Pinnell, 2006) is key to helping children build effective systems for reading and writing. It represents a meta-principle for the instructional system that we sought to measure in this study.

Details about the Developing Language and Literacy Teaching Rubrics

The development of the instrument described in this report drew heavily on prior research and clinical practice within the Literacy Collaborative and the Center for Urban School Improvement at the University of Chicago. Over the last several years, Literacy Collaborative training staffs have used a set of observational protocols to assess classrooms where novice literacy coordinators were in training. These protocols were

based on an earlier small-scale study (Lyons & Pinnell, 2001) that sought to examine the relationship among literacy coaches' practice, the implementation of the comprehensive framework in classrooms, and students' annual learning gains. Based on a long-term set of observations of ten teachers, Lyons and Pinnell developed three scales that assessed teaching and found that student learning gains were greatest in classrooms where teachers scored high on these instruments. While these instruments required further refinement and reliability and validity testing, results indicated that valid assessments could be built on this base.

As noted earlier, we decided to anchor our Developing Language and Literacy Teaching (DLLT) rubrics around the six core instructional components that comprise the framework (see Figure 1).

[[Insert Figure 1, Eight Areas for Rubric Development, about here.]]

We sought to develop two additional rubrics aimed at a more integrative look at literacy instruction:

1. *General Aspects of Teaching* deals with the basic organization of the classroom to support quality instruction. These attributes are crucial to productive activity in any classroom but are especially so in a comprehensive literacy framework where a great deal of activity is occurring simultaneously at any given time. Here, the observer uses the rubric to evaluate classroom materials and organization, student engagement, the quality of teacher-student interactions, and sense of community among students across the entire lesson.

2. *Teaching for Strategies* was designed to further differentiate truly expert practice from more mid-level efforts. The goal of reading instruction is to have students learn ways of thinking: Literal thinking (in fiction and nonfiction texts), inferential and analytic thinking, word solving (including phonics and word analysis), and fluency and phrasing. Likewise, it seeks to help students become fluid, expressive writers who understand how texts are organized, who can logically advance an argument, and who have developed voice. Accomplishing these goals in the classroom involves more than fidelity in implementing specific instructional activities. It also entails coordinating students' learning experiences so that they "add up" to produce strong, self-extending readers and writers.

Using rubrics based on the six core components, as well as the two rubrics outlined above, we sought to capture subtle differences between effective and less effective teaching. We conceptualized these differences as representing a continuum of learning for an individual teacher.

The rubric for Guided Reading consists of seven elements and is presented in Figure 2. Each element (or row in the display) focuses on a specific key aspect of this instructional practice. The ratings in each row move from simple descriptions of base practice at level one to a cumulative description of expert practice at level four. The narrative descriptions are intended to provide a sense of what change in practice looks like for each of these aspects of instruction (rather than giving a simple numeric score).

[[Insert Figure 2, Guided Reading Rubric, about here]]

As examples, we offer an explanation for the scoring of row six of the rubric that focuses on the selection of teaching points after the reading in the small group:

1. A teacher who is just beginning to provide small group instruction (guided reading) might begin by simply going through the steps. The lesson might *look* like guided reading, but the teacher is operating on basically the same theory as previously used while implementing other approaches—in other words, simply taking on superficial characteristics rather than teaching intensively for strategic actions. Thus, after the reading, she may not explicitly respond to what she has observed as students read. The score would be a 1—“Makes no teaching points, even though there were opportunities to do so.”

2. After learning more (possibly through coaching), the same teacher might begin to try to teach for strategies and the score could be a 2—“Makes teaching points, but they do not help students to engage in effective processing of texts.” Here, the teacher is taking on the aspects of the approach but not connecting with students.

3. Over time and with self-analysis, the teacher might score a 3—“Makes teaching points, but not all of the teaching points help students engage in effective processing of text.” Here, the teacher is doing a good approximation of the approach and making a positive impact, but is not always responsive to students’ learning.

4. Finally, at the highest level on the rubric at a score of 4, teaching is marked by—“Makes superbly chosen, specific teaching points that help students engage in effective processing of texts.” At this level of expertise, the teacher actively incorporates observations during the reading and makes explicit teaching points that often draw the students back into the text to illustrate the problem-solving.

The full rubric for guided reading requires the rating of each aspect of instruction represented by the seven rows. The teacher may be more or less proficient on different aspect of the rubric. That is, she could be strong at supporting individual students during reading but relatively weak in selecting and engaging students in explicit teaching points after reading. The end result is a guided reading profile, consisting of multiple scores, for any observed lesson. Drawing on this conceptualization, we sought, in developing the DLLT, to maximize the likelihood that we would be able to capture an accurate sense of teacher learning over time.

Similar logic played out in conceptualizing the developmental rubrics for the remaining five instructional components. The overall goal of each rubric is to distinguish more procedural teaching (or “going through the motions”) from skillful teaching from genuine expertise. In addition, the same teacher would be rated using the last two rubrics on General Aspects of Teaching and Teaching for Strategies, which requires observers to think analytically across contexts at the same point in time.

Taken together, these eight DLLT rubrics seek to examine teachers’ development within the separate components of the instructional framework, as well as their capacity to weave these elements into effective instruction based on their pedagogical knowledge and their ongoing observation of children. Moreover, the rubrics are intentionally not “generic” in orientation. Rather, embedded within them is an explicit theory of student development in reading and writing and a corresponding language and literacy instructional framework.

Field Testing, Reliability, and Validity Analyses

Next, we describe the design for field testing draft rubrics, along with the training protocol, the actual data collected and the analysis of the multiple elements of the rubric as a scale.

Data Collection Design

The field trial for the draft rubrics was carried out in 22 schools in four Ohio school districts. Each selected school had at least one experienced literacy coordinator who had been trained by the Literacy Collaborative at The Ohio State University. The basic design plan called for each coordinator to observe four teachers, one at each grade level K–3, on five occasions (once a month) from January through May of 2005. Coordinators within each school were asked to select two novice teachers (two years or less experience with the Literacy Collaborative framework) and two experienced teachers (three or more years of practice with the framework) to include in the study. We asked the literacy coordinators to observe a full literacy block (between 90 and 150 minutes) and then rate the lesson according to the draft rubrics.

In order to obtain data on inter-rater reliability, a subset of paired observations was designed for both the initial and final time points. On these occasions, a participating coordinator from one school would travel to another school in the same district so that the two coordinators could jointly observe but independently rate the same classroom lesson. In order to minimize the overall data collection burden on the participating coordinators, the paired observations conducted at time points 1 and 5 counted toward their overall commitment to conduct up to twenty observations apiece for our instrumentation study. This resulted in a modification to the basic design plan with some teachers observed on

five occasions, and others on only three occasions (time points 2, 3, and 4) in order to make the gathering of the inter-rater reliability data feasible.

Training Protocol for Using the DLLT Rubrics

The literacy coordinators in the field study were involved in a two-day intensive training session on the use of the rubric. This training involved a structured introduction to each part of the rubric.

1. Researchers described the overall rubric, for example, for a read-aloud, including why particular aspects of the practice was being highlighted for observation while other details may be omitted. Discussion followed about the wording and developmental distinctions that might be observed for the particular practice or rubric element, each a “row” in the overall rubric.

2. The group observed a complete lesson on video and independently rated the lesson using the rubric. Again, observation was followed by discussion in the large group to clarify what they had seen and to reach consensus about the rating.

3. Another video of the same instructional component (enacted by another teacher) was viewed by the group. The discussion that followed was in small group format. Each group talked about their impression of the lesson as a learning opportunity for students.

4. They then looked across each row of the rubric to examine their agreement. In addition, if important aspects of the lesson were discussed that were not in the rubric, these issues were noted.

These small-group discussions were audio and videotaped for further analysis by the researchers (Rodgers & Hung 2006).

Data Collection

We could not fully implement the design as planned in every school for a variety of reasons including the school structure (e.g., only K–2 enrollments), the demography of the school staff (e.g., no novice teachers), the specific assignments of the participating coordinators (not all coordinators work across the K–3 grades), and the refusal of a few individual teachers to participate. In addition, in a small number of cases, teachers who originally agreed to participate subsequently had to drop out, and the full observation protocol could not be completed for them.

In the end, we obtained useable rubrics from 275 literacy block observations. These observations came from seventy-eight classroom teachers in the twenty-two schools. Of this group, thirty-seven teachers were novices within the literacy framework, and forty-one were more experienced. The average coordinator observed 3.5 teachers and the average teacher was observed on 3.5 occasions. In addition, paired observations were conducted on thirty-three lessons at time points 1 and 5, respectively. Each coordinator was involved in at least one joint observation at both of these time points.

Using an Item Response Theory Model to Examine Developmental Scale Properties

Each rubric, including the six content-based rubrics and the two cross-observation measures, consists of multiple elements (ranging from 3 to 10, depending on the

particular rubric domain). Each element is independently rated on a four-point descriptive scale.

A primary analytic task for the pilot study was to examine whether these multiple elements combined together into a theoretically interpretable and empirically defensible developmental scale. The results from the Rasch Rating Scale analyses described below provide key evidence in this regard.

The Rasch Rating Scale model (Wright & Masters, 1982) is an extension of an item response latent trait model that is now commonly employed in standardized test construction. Instead of the dichotomous (correct/incorrect) responses in an IRT model, the Rating Scale model is designed for ordered categorical data as in our four-point developmental rubrics. Our Rating Scale analysis defines a measurement scale that is anchored in the relative probability of a teacher's practice being placed in each developmental category within each element that comprises the rubric. In a properly fitting scale, the rubric elements and subcategories hierarchically arrange in developmental order. Teachers are subsequently "measured" on this scale based on the practice ratings recorded for each of them. The scale units are measured in logits (i.e., the log odds of being in a particular development category on a given element relative to base state practice).

A Rating Scale analysis produces a diverse array of statistics for examining the quality of the underlying measurement system.

- First, there are element difficulty statistics, which estimate the likelihood that a teacher will score highly on that particular rubric element. Pedagogical skills that are relatively easy for teachers to acquire, such as selecting engaging and appropriate books in the

Read Aloud rubric, tend to have lower difficulty estimates than more advanced skills, such as facilitating strong discussion during the reading of the book. Thus, how the element difficulties locate themselves within their respective scales provides critical evidence as to whether the rubric elements are empirically sequencing in a theoretically reasonable order.

- Second, and also important, are a set of element infit statistics. These estimates provide information about the degree to which the teacher ratings on a particular element are consistent with the element's placement in a hierarchically ordered scale. If the DLLT rubrics are truly developmental scales, individuals who are rated high on a particular practice should be more likely to be rated high on the "easier to learn skills" that are below it in the scale and be less likely to demonstrate competence on the "more difficult" to master elements that appear above it in the scale. In essence, the infit statistic measures the degree to which the element's behavior deviates from what we would expect in a perfect developmental scale. Under the Rasch Rating Scale model, a probability value can be attached to the estimated infit statistic associated with each element within a rubric. The underlying null hypothesis associated with this test statistic is of the form "assuming that this element is part of a hierarchically ordered scale, how likely is it that we could get an infit statistic this large by chance alone?" Thus, this aspect of the Rasch analysis provides critical information for judging whether it is appropriate to view the combination of elements in a rubric measure as forming a coherent developmental scale.
- Third, the Rating Scale analysis provides a standard set of statistics about the internal consistency of each scale. The "person reliability statistic" generated in the Rating Scale analysis is an internal consistency measure similar to Cronbach's alpha.

Evidence on the Reliability of the Rubrics

Below we present inter-rater reliabilities of the rubric elements as well as internal consistency reliability and inter-rater reliability for the final rubric measures.

Inter-rater reliabilities of the rubric elements. Obtaining good inter-rater reliability with observational rubrics can be a nettlesome problem, especially when the rubrics are being used by school-based staff (rather than trained researchers) under ordinary field conditions. In response to this concern, we sought to define each element within each rubric as representing a pedagogical practice that was directly observable. We also aimed for clear descriptive language in defining each of the developmental steps within each observation element. We then used the paired observations at time point 1 to help us discern how well our first draft descriptions actually worked when employed by different literacy coordinators. Table 1 presents the results from these observations. We report here on both the exact agreement and adjacent category agreement between raters (i.e., ratings within one category of each other). We initially performed this analysis separately for each element in each rubric. We then averaged across the elements that comprise each rubric to create rubric-level agreement statistics reported in the table.

The inter-rater reliabilities from time point 1 helped us to improve the draft rubric specifications. While most of the initial rubric elements appeared to work well, we made nine changes in the pilot instrument based on results from the time point 1 analysis. Specifically, we dropped one element each from shared reading and guided reading where inter-rater reliability seemed problematic. We clarified the descriptions on four other elements (one in Guided Reading, two in Word Study, and two in Teaching for

Strategies), and we added one entirely new element to the Teaching for Strategies rubric. This revised set of scales was then used in the observations at time points 2 through 5. We also added a component to our rubric training protocol where an experienced staff member from the Literacy Collaborative at The Ohio State University jointly observed a lesson with each literacy coordinator at their own school. The staff members independently rated the lesson and then discussed their differences, if any, in the use of the rubrics.

Taken together, these changes resulted in a substantial improvement in the inter-rater agreements element by element from time points 1 to 5. On every rubric, the average adjacent category agreement now exceeded 90 percent. The exact agreements, while lower, improved substantially and hovered around 60 percent for all of the scales.

[[Insert Table 1, Inter-rater Agreement for the DLLT Rubrics, about here.]]

Finally, we also computed a weighted Kappa statistic (Fleiss 1981; Landis & Koch, 1977) on the time point 5 data for each element in each rubric (see Table 2). This statistic provides the most detailed and accurate assessment of agreement in ordered categorical data, which is the type used in our rubrics. The statistic combines information on both agreements and *the degree of disagreements* to produce an overall index of inter-rater reliability. A Kappa ratio of 0.80 indicates outstanding agreement while values from 0.60 to 0.79 represent substantial agreement. Values from 0.40 to 0.59 are considered moderate inter-rater reliability.

In terms of the DLLT, 20 individual rubric elements had weighted Kappas of 0.70 or higher. The weighted Kappas were less than 0.50 on only six elements and between 0.50 and 0.60 on another 13 elements. Overall, this suggests quite good field agreement element by element within each rubric. We undertook one final review of the six elements with Kappas less than 0.50, but found no reason to introduce any further changes in any of these.

[[Insert Table 2, Weighted Kappa Statistics for Each DLLT Rubric Element (Time Point 5), about here.]]

Internal consistency reliability for the final rubric measures. Table 3 provides reliability estimates from the Rasch Rating Scale analysis on the eight separate rubrics that compose the DLLT. Six of the eight rubrics achieved internal reliability of 0.75 or higher. The reliability for the Read Aloud and Shared Reading rubrics were weaker, at 0.63. This is not surprising, given that these last rubrics consist of only three rating elements apiece. As expected, the estimated reliability was higher for those rubrics with a larger number of rating elements. For example, Teaching for Strategies, which consists of ten rating elements, has an internal consistency reliability of 0.91.

[[Insert Table 3, Internal Reliability of the DLLT Rubrics, about here.]]

Our Rating Scale analysis suggests that each of the rubrics has good internal consistency and that most of these scales, with the exception of Read Aloud and Shared Reading, can be used separately in subsequent analyses.

Inter-rater reliability for the final rubric measures. Finally, we computed correlations for the eight rubric measures across the thirty-three paired observations conducted at time point 5. These are reported in Table 4. All of these correlations are 0.80 or higher, except for Shared Reading, which was 0.78. These results indicate quite good inter-rater agreement at the rubric measure level. These are key reliability statistics for our use of the DLLT in the context of the IES study. They indicate that two different observers, using the same observation protocol on the same lesson, will produce good score agreement at the rubric measure level. This is very encouraging because all our subsequent analyses for the larger four-year study of change in practice will involve use of these measures.

[[Insert Table 4, Correlations among Measures for Paired Observations at Time Point 5, about here.]]

Evidence for Examining Construct Validity

As noted earlier, the Rasch analysis also provides evidence for examining the construct validity of our rubrics as developmental measures of teaching practice. Specifically, scrutiny of the item-difficulty statistics allows us to investigate whether the empirical ordering of the items follows the theoretically expected development pattern within each rubric. Practices that teachers find easy to integrate into their classrooms should have low

item difficulty. Those items that are at the higher end of the difficulty scale should represent more expert practices that are more difficult to instantiate in classrooms and as a result are less frequently observed. Thus, by examining the ordering of the estimated difficulties within each rubric we have a face validity test of the scale as a developmental rubric. Quite simply, do the difficulty estimates form an interpretable developmental sequence?

Also of value are the misfit statistics that evaluate the consistency of the item map across teachers. If the scale represents a general developmental profile that most teachers follow, we would expect small item-misfit statistics. If different teachers take on aspects of the practice in different orders, however, the misfit statistics will be large, indicating that the elements of the rubric do not cohere as an overall developmental measure. Taken together with the item difficulties, the associated misfit statistics provide empirical evidence to examine our theoretical assumptions about how teachers move from less to more expert practice and what this means substantively. We detail below our analysis of the results for Guided Reading, Writing Workshop, and Word Study. The remaining difficulty maps can be found in Appendix A.

Guided Reading. Figure 2 displays this item difficulty map. Each bullet-pointed description represents an element (i.e., a row) of the rubric. The easiest item, at the bottom of the map, is “selection of an appropriate text for the reading level of the group.” With appropriate teacher guidance, such a text has potential for significant learning opportunities for students. This is the most basic routine to be established in a classroom where Guided Reading is being introduced. Considerably more difficult for a teacher is the preparation of a book “introduction to engage students in a discussion around

meaning.” This practice is key in preparing students to work with ideas they will encounter in the text that are challenging or new to them. A good book introduction supports students in conversation that draws on their prior knowledge and facilitates their access to new information. It can also involve the integration of talk about potentially demanding characteristics of the book and sets the stage for successful problem solving of vocabulary, figurative language, or other aspects particular to this text.

[[Insert Figure 2, Guided Reading Item Map, about here.]]

Moving a bit further up the scale, the difficulty map indicates that teachers who are providing strong introductions also are likely to be prompting for effective reading strategies as they listen to student reading. (Notice that the clustering of these three item difficulties within 0.5 logits of one another.) Conceptually, this makes sense. Teachers who provide a supportive book introduction are also likely to prompt students as they listen to individual students read the text. The same strategic thinking that teachers display in the book introduction also manifests itself in response to students’ reading.

After the guided reading lessons, expert teachers pull together what they have seen in student reading behaviors in an effort to solidify and reinforce their students’ learning. As evidenced by the item difficulties, these aspects of guided reading—both extending meaning through student talk and especially making explicit teaching points—are less frequently observed. This is not surprising, in that this instructional practice entails in-the-moment decision making about what will most effectively demonstrate to students what they have learned to apply and/or how they can further extend their

problem solving skills. Only expert teachers who are keen observers and interpreters of student reading behaviors are likely to demonstrate this practice. Moreover, for powerful instruction to occur, the teaching point has not only to be well made, but it also needs to be responsive to where students are as developing readers.

Overall, the Guided Reading item map presents a coherent progression from basic routines toward aspects of practice that require more processing of information by the teacher, both in terms of the demands of the text and the reading behaviors of students. Moreover, because we found no evidence of large misfit statistics, the scale appears to define a general progression that applies for most teachers.¹ Taken together, this evidence suggests that it is appropriate to view this rubric as a developmental scale that moves from novice to more expert practice.

Writing Workshop. One of the first aspects for teachers to take on within this framework component is “organizing the class so there is time for individual conferences with students about their work.” While these conferences may be readily introduced into classrooms, they can be more or less productive for students, based on what occurs in the particular conference as well as the rigor of the other instructional supports and activities the teacher provides for students when they are waiting for their individual conference. The next cluster of items on the difficulty map focuses on the minilessons conducted as part of the writing workshop. The minilesson provides opportunities to teach specific writing principles. The most basic aspect of the minilesson is the selection and “demonstration of an example of the writer’s craft.” This explicit teaching of a core writing principle supports students in analyzing what they read to inform what and how

¹ A table of the complete item difficulties and fit statistics appears in the appendix.

they communicate in their own writing. More proficient teachers go a bit further, to “carefully check students’ understandings” of the connections of the principle articulated in the minilesson to their own work. For example, the teacher may ask students to discuss a selected piece of writing in pairs while she circulates to participate in these conversations. Finally, each minilesson sits within a broader instructional agenda in which teachers seek to help students come to understand the processes of good writing. Relating the minilessons to these broader writing principles provides multiple occasions for helping students to develop the cognitive schema necessary to scaffold their increasingly independent work as writers. It is this orchestration of the minilesson as a whole that constitutes the most powerful instruction.

As was true for guided reading, we found no evidence of item misfit in this scale either. (See Appendix A for details.) As a result, we can again interpret the ordering of the element difficulties as representing a developmental scale. This means, for example, that teachers whose minilessons are expert also are likely to conduct strong individual conferences. In contrast, these same teachers may not yet engage general classroom conversations classroom around student writing that push the writing process forward. This expertise is captured in the last two elements that anchor the top of this measurement scale.

[[Insert Figure 3, Writing Workshop Item Map, about here.]]

The most difficult items within the measure map for the Writing Workshop rubric involve the effective use of student sharing of their writing as opportunities for explicit

teaching. This sharing of writing creates spontaneous “teachable moments” that can help make the essence of good writing concrete both for the students who are sharing and for others in the class. Such instruction makes demands on teachers to interweave explicit points that return students to the writing principles mentioned earlier and, in essence, close the loop of the entire activity. It makes demands on teachers to integrate what they may see in a specific piece of a student’s writing and what they know through observation of this student’s writing behaviors over time. Finally, at the most expert level (i.e., the top of the item difficulty map), teachers are skillful in facilitating student comments on each other’s writing. This level of expertise is evidenced in classrooms where students, in talking about their own work, display an understanding of the writing process and use these ideas to scaffold their responses to other students’ writing as well. To be sure, such talk does not occur spontaneously. It must be modeled and facilitated by a skillful teacher.

Taken overall, the item map for Writing Workshops explicates a coherent developmental set of practices. It moves from implementing a structure of activities including conferencing, minilesson, and sharing to a sophisticated crafting of explicit instructional points in response to what the teachers observe as students engage in authentic writing. As with Guided Reading, the most demanding aspects for teachers are the “in-the-moment decisions” that attempt to connect and integrate students’ writing experiences with the core principles of the writing process. At base, here, is an effort to foster students’ development of an overall schema to guide their subsequent work as writers. The item difficulty map suggests that such instruction is the hallmark of teacher expertise in the Writing Workshop.

Word Study. This item map is displayed in Figure 4. The easiest items involve selection of a task that students can engage independently and that has potential for student learning around how words work. This routine involves marshalling appropriate materials to set the context for learning. Somewhat more difficult than selecting an appropriate task is using examples in the minilesson so they create a clearly understood link between the task and the word study principle that is the focus of the lesson. Of course, examples can be used simply to show students primarily how “to do” the activity that will follow. Elaborating the word study principle as a clear and explicit part of the minilesson is a more sophisticated aspect of teaching within Word Study, and this is reflected in its placement on the item difficulty map.

Demonstration of how the principle relates to the application task is still another dimension for teachers to attend to in the minilesson. The teacher orchestrates links so that students understand their work, for example, not simply as sorting words that have the same ending pattern but as a task that helps them understand how letter-sound relationships work (and how they can use this in their reading).

[[Insert Figure 4, Word Study Item Map, about here.]]

The most expert aspects of Word Study are reflected in the rubric elements around how teachers use these sharing opportunities to advance student learning. Similar to Writing Workshop, this sharing activity provides a context to reinforce core principles about how words work. The teacher does this both by restating the principle based on student work and facilitating conversations in which students reflect on what they have

learned in this regard. These two aspects of Word Study instruction are the two most difficult items in the measure.

So, overall, these results for the Word Study scale rank order exactly as we would expect given base pedagogic theory. In addition, we found no evidence here, either, of item misfit in this scale. This evidence is consistent with the proposition that the Word Study scale function as a valid metric for charting teacher development.

*Evidence Concerning the Usefulness of the Rubrics for Capturing Teacher Learning
Over Time*

These analyses offer the opportunity to discuss a developmental model of teacher practice and to examine measured difference in classroom practice between novice and experience teachers.

A Developmental Model of Teacher Practice. Looking across the item maps of the elements of the framework, a pattern of teacher development begins to surface—not just about how teachers implement an instructional activity, but also about the developmental path teachers traverse as they become more expert practitioners of these activities to support student learning.² Figure 5 summarizes the grounded model that emerges from the empirical results.

² We undertook the same conceptual analysis for each of the other three instructional components in the Comprehensive Literacy Framework (Read Alouds, Shared Reading, and Interactive Writing) and the two general rubrics (General Aspects of Teaching and Teaching for Strategies). In general, the item difficulties map along similar lines to those detailed above. Moreover, we found very little evidence of item misfit (see Appendix A for details) on any of the eight rubrics. This finding suggests that it is appropriate to interpret

[[Insert Figure 5, Developmental Model of Teacher Practice, about here.]]

Initially, teachers focus on the organization of the practice as well as the materials and activities that are necessary to implement it in their classroom. They move toward embedding the principles of the reading process within this structure to support student learning. This progression occurs in minilessons and other aspects of instruction and increasingly becomes a guide to teachers' selection of materials and texts to engage the whole class and small groups. As teachers become more adept at demonstrating these principles through examples, their observation of students becomes even more critical. It is their combined understanding of the reading process and observations of how their students engage in problem solving and thinking about texts that allows teachers to identify and create more explicit teaching points in direct response to student learning.

Effectively linking knowledge and observation represents a significant jump in teacher practice as heuristically displayed by the difficulty gap in Figure 5. Teachers operating at this level are much more strategic and targeted in their selection of the principles they introduce and which they repeatedly return to. Teachers are now also building opportunities for students to articulate and share what they are learning. This facilitation of student talk during sharing and discussion reinforces student learning and creates a self-extending quality to the practice.

the results as representing a set of general processes of developing expertise within a Comprehensive Literacy Instructional framework.

Taken together, the rubrics assess a developmental pathway toward more expert practice as teachers respond to where students are as learners. In this regard, the rubrics represent a more complex endeavor than might be implied by a simple measure of fidelity of program implementation. The rubrics do assess whether appropriate classroom structures are in place and teachers' automaticity in executing basic instructional routines. Both of these are essential bases for good practice. Expert instruction in comprehensive literacy, however, requires considerably more than this. Teachers must integrate the use of these structures and routines around their research-based understandings of the reading and writing process; with their specific knowledge of the developing literacy skills for each child; and aiming to accelerate the progress of all children toward the ultimate goal of becoming effective, independent readers and writers.

Comparison of novice and experienced teachers. The DLLT rubrics have been specifically designed to assess teacher development over time in the pedagogical practices of a comprehensive literacy framework. A critical test of their validity is whether teachers actually demonstrate change on these measures as they continue to be exposed to LC professional development. This question is under investigation now in our longitudinal field study.

As part of the basic design for the instrumentation study, we sampled participants according to the number of years that each teacher had been exposed to Literacy Collaborative training. By design, approximately half of the teachers were *new* to Literacy Collaborative (less than two years of professional development) and the other half more *experienced* (two or more years in the program). According to the basic program description offered by the Literacy Collaborative, it takes three years of

professional development support for a typical teacher to learn to work with some proficiency in the framework. This suggests that we should see significantly different reports on the DLLT rubrics if we subdivide the sample by their experience level. Figure 7 presents a set of box plots comparing these two groups on the eight separate measures. (Note that all of the mean differences presented here are statistically significant beyond the 0.001 level.)

In general, we found large differences on the rubrics that assess teacher development on the six core instructional components of a comprehensive literacy framework. For example, the score for a median *experienced* teacher in Guided Reading and Writing Workshop is equivalent to about the 75th percentile in the distribution for *new* teachers. Similar size differences exist for the overall rubric, General Aspects of Teaching rubric.

Perhaps more complex than becoming expert in any one instructional component is marshalling these activities so that the reading and writing process become more coherent for students across these contexts. An even larger difference appears when we compare *new* and *experienced* teachers on Teaching for Strategies. On this rubric, the 25th percentile *experienced* teacher is scoring at about the same level as a 75th percentile *new* teacher. Recall that this rubric draws on evidence from the entire literacy block (rather than just when a particular component is being taught) and seeks to assess how well teachers integrate their pedagogical efforts across all elements of the framework toward strategic teaching to advance students literacy learning. In a sense, this is our single most direct measure of expert teaching, and not surprisingly, it discriminates the most across the experience dimension.

Implications

In developing the rubrics detailed in this paper, we drew on empirical research in elementary literacy learning and clinical expertise regarding literacy instruction and professional development to advance this instruction. Our research team combined a diverse collegueship of expertise, including extensive experience observing teacher practice, coaching for teacher development, literacy subject matter knowledge, psychometrics, and research methodology. Each of these fields played an essential role in building, refining, and field testing the rubrics for teacher development presented here.

Our rubrics direct attention to both teachers' instructional behaviors and language use in the classroom within the context of a *particular instructional system* (albeit a fairly general one of comprehensive literacy). This instructional system context afforded us considerable specificity in defining the observation protocol, as it allows us to target (and evaluate against) a set of structures and routines that *should* exist in every classroom. Similarly, the extensive clinical observations that have now accumulated about comprehensive literacy practices guided us in conceptualizing the variations in practice that were likely to occur. As a result, we were able to design a set of rubrics where both high levels of both inter-rater and over-time reliability were relatively easy to achieve. In this latter regard, we were mindful of recent findings from Rowan, Camburn, and Correnti (2004), who found that fifteen or more observations were needed to assess reliably teacher practice. We were relieved to learn in our study that it is possible to achieve reliable and valid measures of teaching practice with as few as three observations per year. In the broadest sense, the ability to draw on an empirically grounded framework

for literacy practice reduced the burden placed on assessment system design that typically seek to obtain reliable and valid measures that might work in any context under any system of instruction.

In conclusion, we offer three somewhat more general comments about the implication of our research for future efforts in this general domain.

Research on the Causal Cascade Connecting Professional Development to Student Learning

The work on measuring teacher development detailed in this article sits within a larger research context of the links between a professional education initiative and improvements in student learning. Any integrated research program of this sort must examine a set of interrelated questions that has the character of a causal cascade. This cascade begins with the intentional design of a professional development program. Here, we ask: “What exactly do we expect teachers to learn and how would we know if they learned it?” Next, assuming that some knowledge and skill has been acquired, we proceed to a second question: “Does this new knowledge and skill actually translate into observable changes in classroom practice, and how would we know this? Third, assuming positive evidence with regard to questions one and two, we proceed to ask: “Do these developing classroom practices actually lead to improvements in student learning as assessed through reliable and valid measures of this learning?” This should be the basic organizing paradigm for research on the effects of professional education.

Unfortunately, in many efforts to promote changes in teacher practice, clear specification of the changes expected in practice, as well as a system for their reliable and

valid measurement, are often neglected. In essence, this creates a huge “black box” at the heart of the intended change process. Enormous energy and resources may be placed into the design and implementation of professional education that then lacks the ability to actually track the mechanisms through which this is intended to improve student learning. Absent this, we don’t know how to interpret fully the observed student outcomes and we lack empirical evidence for formative evaluation and subsequent fine-tuning of the professional education initiatives.

In our continuing research with these rubrics, we will directly examine the link, if any, between teachers’ exposure to professional education and the development of practice in their classroom as measured on these rubrics. We also will seek to evaluate how measured changes in classroom practice relate to changes in student learning in these same classrooms over time. If these empirical links are subsequently validated, then the rubrics would ultimately provide a key indicator both for judging the efficacy of professional development initiatives and for predicting the likely improvements in student learning that might follow. In our view, developing reliable and valid tools of this sort is an essential step in seeking to advance more ambitious instructional efforts, like comprehensive literacy, at scale.

Advancing Literacy Practice Improvement at Scale

The catalyst for developing these rubrics was to provide instrumentation for a research study, but it is important to recall that our data were collected by reading coaches as a regular part of their work. This indicates that this tool can be used reliably by coaches in the clinical context of everyday school practice. This is important because the “common

language” embedded in “common tools” may be a key, perhaps even an essential resource, to promoting individual learning and advancing practice improvement at scale. One of our next steps will be to adapt these rubrics for use by coaches as a vehicle for guiding their efforts to fostering learning communities in their schools. Differentiated staff development for novice and more experienced teachers is almost nonexistent in today’s professional development. We envision coaches using the rubrics, for example, as a lens to focus their observations in classrooms, making it possible to create more targeted individual professional development plans for teachers, and in identifying larger, school-wide professional development priorities. We also envision the rubrics as a tool for use by teachers to guide their own self-reflection.

Perhaps most significant, the DLLT rubric represents a critical reframing of how we think about changes in teacher practice as a result of intentionally designed professional development. Most large-scale instructional reform initiatives tend to view teacher change through a *fidelity of implementation lens*. This view directs attention to whether teachers are using the materials and activities in the ways intended by the program designers. This is, of course, not irrelevant to a comprehensive literacy program. However, the deeper goal in such programs is to develop teachers’ capacity as reflective practitioners. Expert practice entails attention to where their students are as readers and writers, cognizance of the next “goals in view” for student development and the skillful organizing and execution of appropriate instruction in response.

In essence, the DLLT rubrics offer an *alternative lens*—one that focuses on *teachers’ development of expertise within an instructional system*. We conceive of the latter as moving from the introduction and experimentation with new methods in

classroom practice to achieving procedural automaticity to evidence of genuine expertise in the orchestration of their use to advance student learning. We believe that this affords a powerful heuristic for thinking about the problems of improving complex instructional practice at scale—and one that teachers will likely readily embrace as a valid and respectful accounting of their own continuous improvement efforts.

Nudging the Methods for Rubric Development Forward

Finally, the DLLT rubrics also introduce a new measurement model strategy that shows promise for improving future rubric construction for charting teacher development. Clinical rubrics tend to strive for a rich description of instruction by combining multiple aspects of classroom practice. Using such descriptive rubrics, however, it can be difficult to achieve traditional standards of reliability and validity. Individual teachers may develop unevenly on the different components that comprise these composite descriptions, thereby forcing observers to make a somewhat arbitrary choice as to which category value to assign. The DLLT, in contrast, by treating each row of the rubric as in essence a rating scale element unto itself, avoids much of this difficulty. Rich composite descriptions of practice still result, but how the elements combine in a descriptive account of practice for a particular teacher at a specific time is now allowed to vary. In addition, the IRT Rasch analysis permits a careful examination of the interrelationships among these descriptive elements (i.e., information contained in the multiple rows that describe teacher development within a specific pedagogical practice such as guided reading). This affords valuable information for assessing construct validity while simultaneously strengthening instrument reliability.

The work described here has been supported by a Teacher Quality Grant from the Institute for Educational Sciences, R305M040086, “Can Literacy Professional Development be improved with Web-Based Collaborative Learning tools?: A Randomized Field Trial.” We are appreciative of this support from IES. All errors of fact, omission and/or interpretation are solely the responsibility of the authors.

References

- Beck, I. L., & McKeown, M. G. (2001). Text talk: Capturing the benefits of read-aloud experience for young children. *Reading Teachers*, 55, 10–20.
- Clay, M.M. (2001). *Change over time in children’s literacy development*. Portsmouth, NH: Heinemann.
- Darling-Hammond, L. (1996). What matters most: A competent teacher for every child. *Phi Delta Kappan*, 78, 193–200.
- Darling-Hammond, L., & McLaughlin, M. W. (1996). Policies that support professional development in an era of reform. In M. McLaughlin & I. Oberman (Eds.), *Teacher Learning: New Policies, New Practices* (pp. 202–18). New York: Teachers College Press.

- Englert, C. S. (1984). Measuring teacher effectiveness from the teacher's point of view. *Focus on Exceptional Children, 17*, 1–15.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.
- Fountas, I. C., & Pinnell, G. S. (2006). Teaching for comprehension and fluency: Thinking, talking, and writing about reading, K–8. Portsmouth, NH: Heinemann.
- Gersten, R., Baker, S. K., Haager, D., & Graves, A. W. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners: An observational study. *Remedial and Special Education, 26*, 197–206.
- Graves, A. W., Gersten, R., & Haager, D. (2004). Literacy instruction in multiple-language first-grade classrooms: Linking student outcomes to observed instructional practice. *Learning Disabilities Research & Practice, 19*, 262–272.
- Haager, D., Gersten, R., Baker, S., & Graves, A. (2003). The English language learner observation instrument for beginning readers. In S. Vaughn & K Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning*. Baltimore, MD: Brookes.
- Junker, B., & Matsumura, L. C. (2006). Beyond summative evaluation: The Instructional Quality Assessment as a professional development tool (CSE Technical Report 691). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.
- Junker, B., Weisberg, Y., Matsumura, L. C., Crosson, A., Wolf, M., Levison, A., & Resnick, L. (2005). Overview of the instructional quality assessment (CSE Tech.

- Rep. No. 671). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.
- Landis, J. R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.
- Lundberg, I., Frost, J., & Peterson, O. P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly*, *23*, 264–84.
- Lyons, C. A., & Pinnell, G. S. (2001). *Systems for change in literacy education: A guide to professional development*. Portsmouth, NH: Heinemann.
- McCarrier, M. C., Pinnell, G. S., & Fountas, I. C. (2000). *Interactive writing: How language and literacy come together*. Portsmouth, NH: Heinemann.
- National Institute of Child Health and Human Development. (2001). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Washington, DC: U.S. Department of Health and Human Services, NIH Pub. No 00-4754.
- Pearson, P. D., & Fielding, L. (1991). Comprehension Instruction. In R. Barr, M. Kamil, P. Mosenthal, & P. D. Person (Eds.), *Handbook of reading research* (vol. 2, pp. 815–60). New York: Longman.
- Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, R. B., & Beatty, A. S. (1995). *Listening to children read aloud: Data from NAWP's Integrated Reading Performance record at grade 4* (Report No. 23-FR-04). Prepared by Educational Testing Service under contract with the National Center for

- Education Statistics, Office of Educational research and Improvement, U.S. Department of Education.
- Pinnell, G. S., & Fountas, I. C. (1998). *Word matters: Teaching phonics and spelling in the reading/writing classroom*. Portsmouth, NH: Heinemann.
- Pressley, M. (1998). *Reading instruction that works: The case for balanced teaching*. New York: Guilford Press.
- Resnick, L., & Junker, B. (2006). *Using the instructional quality assessment toolkit to investigate the quality of reading comprehension assignments and student work* (CSE Report 669). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.
- Rodgers, E., & Hung, C. (2006). Understanding the work of coaching: A lens for viewing classroom practice. Paper presented at the American Education Research Association, San Francisco.
- Schon, D. (1993). *The reflective practitioner*. New York: Basic Books.
- Snow, C. E. (2002). Reading for understanding: Toward a research and development program in reading comprehension. Prepared for the Office of Research and Improvement. Santa Monica, CA: RAND Corporation.
- Snow, C., Burns, M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Stanovich, P. J., & Jordan, A. (1998). Canadian teachers' and principals' beliefs about inclusive education as predictors of effective teaching in heterogeneous classrooms. *Elementary School Journal*, 98, 221–38.

Sterbinsky, A., & Ross, S. M. (2003). *Literacy observation tool reliability study*.

Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Vellutino, F. R., & Scanlon, D. B. (1987). Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study. *Merrill Palmer Quarterly*, 33, 321–63.

Wright, B. D., & Master, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Figure 1

Eight Curriculum Areas for Rubric Development	
<i>Rubrics Related to Specific Instructional Contexts</i>	
1. Interactive Read Aloud	Usually whole class instruction.
2. Shared Reading	Whole class or small group instruction.
3. Guided Reading	Small heterogeneous groups.
4. Interactive Writing	Whole class or small group instruction.
5. Writing Workshop	Whole class lessons and individual conferences
6. Word Study	Usually whole class instruction, followed by individual application.
<i>Rubrics Used Across Instructional Contexts</i>	
7. General Aspects of Teaching	
8. Teaching for Strategies across Contexts	

Guided Reading — Group 1 Time began: _____ Time ended: _____ *Element not present during the observation*

Text Selection: The teacher:

<p>1 __ Selects a text that is not the appropriate level for the group.</p>	<p>2 __ Selects a text that is the appropriate level for the group, but provides few opportunities for students to learn.</p>	<p>3 __ Selects a text that is the appropriate level for the group and provides some opportunities for students to learn.</p>	<p>4 __ Selects a text that is the appropriate level for the group and is very well matched to the group's needs and provides many opportunities to learn.</p>
---	---	---	--

Text Introduction: The teacher:

<p>1 __ Provides for some introductory activities that may be present, but does not attend to the central elements of an introduction (meaning of whole text, language, aspects of print)</p>	<p>2 __ Provides an introduction that includes some or even all elements (meaning of whole text, language, aspects of print), but is fragmented and not cohesive.</p>	<p>3 __ Provides an introduction that includes some or all elements (meaning of whole text, language, aspects of print), but is somewhat uneven.</p>	<p>4 __ Provides an introduction that includes some or all elements (meaning of whole text, language, aspects of print) in a highly integrated, engaging, and cohesive way.</p>
<p>__ Does not engage children with the text or in interaction with the teacher or other students.</p>	<p>__ May engage children in some conversation, but talk is unfocused and does not help them engage with meaning of the text.</p>	<p>__ Engages children in conversation; some of the talk helps them engage with the meaning of text.</p>	<p>__ Engages students in a conversation that brings them into the text and supports their thinking about the meaning of the text.</p>

During Reading: The teacher: If teacher has appropriate reasons for simply listening to oral reading or letting children read on their own, check this box and record no rating for this row.

<p>1 __ Either does not sample oral reading or interrupts too much with interactions that take the reader “off track.”</p>	<p>2 __ Samples oral reading; interactions give children “clues” for guessing or tells words, but provides little help in engaging in effective reading behaviors.</p>	<p>3 __ Samples oral reading and provides some demonstrations and sometimes prompts for (as needed) effective reading behaviors.</p>	<p>4 __ Samples oral reading and demonstrates, reinforces, and consistently prompts (as needed) for effective reading behaviors and problem solving actions.</p>
--	--	--	--

<i>After Reading: The teacher:</i>			
1 __ Does not engage children in discussion of the meaning of the text.	2 __ Engages children in discussion after reading but talk is unfocused or sometimes off topic.	3 __ Engages in some discussion of the meaning of the text. Students make comments that indicate they are thinking about the meaning of the text.	4 __ Engages children in a rich discussion of the meaning of the text that is evident in their comments about their thinking.
__ Makes no teaching points, even though there were opportunities to do so.	__ Makes teaching points, but they do not help students to engage in effective processing of texts.	__ Makes teaching points, but not all of the teaching points help students engage in effective processing of text.	__ Makes superbly chosen, specific teaching points that help students engage in effective processing of text.
<i>Word Work: The teacher:</i>			
<input type="checkbox"/> Optional: Check and do not record rating if not present during the observation.			
1 __ Shows something about words, but the work is either too easy or too hard for students and may interfere with learning. Word work may involve teaching words “from the book.”	2 __ Shows something about words, but the teaching is not specific and clear, and there is no evidence that students understand the task.	3 __ Shows children something about words. Students participate and perform the task with some understanding.	4 __ Shows children something explicit and strategic about how words work. Students are engaged and there is evidence that they are learning more about word solving.

Figure 2. Guided Reading Rubric

Table 1. Inter-Rater Agreement for the DLLT Rubrics

	<u>Read Aloud</u>		<u>Shared Reading</u>		<u>Guided Reading</u>	
	<i>Obs 1</i>	<i>Obs 5</i>	<i>Obs 1</i>	<i>Obs 5</i>	<i>Obs 1</i>	<i>Obs 5</i>
Exact Agreement	53%	74%	36%	59%	58%	71%
Adjacent Agreement	90%	98%	84%	95%	92%	92%

	<u>Interactive Writing</u>		<u>Writing Workshop</u>		<u>Word Study</u>	
	<i>Obs 1</i>	<i>Obs 5</i>	<i>Obs 1</i>	<i>Obs 5</i>	<i>Obs 1</i>	<i>Obs 5</i>
Exact Agreement	53%	53%	63%	69%	64%	71%
Adjacent Agreement	87%	93%	90%	93%	91%	92%

	<u>General Aspect of Teaching</u>		<u>Teaching for Strategies</u>	
	<i>Obs 1</i>	<i>Obs 5</i>	<i>Obs 1</i>	<i>Obs 5</i>
Exact Agreement	49%	58%	42%	57%
Adjacent agreement	94%	95%	88%	92%

Table 2. Weighted Kappa Statistics for Each DLLT Rubric Element (time point 5)

<i>Read Aloud</i>		<i>Writing Workshop</i>		<i>General Aspects of Teaching</i>	
RA1	0.77	WW1	0.83	GAT1	0.62
RA2	0.76	WW2	0.62	GAT2	0.74
RA3	0.96	WW3	0.43	GAT3	0.45
		WW4	0.73	GAT4	0.51
		WW5	0.86	GAT5	0.50
<i>Shared Reading</i>		WW6	0.82	GAT6	0.72
SR1	0.64	WW7	0.78	GAT7	0.53
SR2	0.71			GAT8	0.66
SR3	0.51			GAT9	0.58
		<i>Word Study</i>			
<i>Interactive Writing</i>		WS1	0.84		
IW1	0.58	WS2	0.67	<i>Teaching for Strategies</i>	
IW2	0.40	WS3	0.51	TS1	0.62
IW3	0.92	WS4	0.62	TS2	0.52
IW4	0.61	WS5	0.54	TS3	0.37
IW5	0.42	WS6	0.85	TS4	0.76
		WS7	0.74	TS5	0.56
				TS6	0.56
		<i>Guided Reading</i>		TS7	0.58
		GR1	0.65	TS8	0.36
		GR2	0.59	TS9	0.72
		GR3	0.75	TS10	0.58
		GR4	0.68		
		GR5	0.78		
		GR6	0.68		
		GR7	0.74		

Table 3. Internal Reliability of the DLLT Rubrics

Rubric	Person Reliability	Number of Elements
Read Aloud (n=234)*	0.63	3
Shared Reading (n=120)	0.63	3
Guided Reading (n=218)	0.79	7
Interactive Writing (n=104)	0.75	5
Writing Workshop (n=213)	0.80	7
Word Study (n=148)	0.83	7
General Aspects of Teaching (n=325)	0.90	9
Teaching for Strategies (n=323)	0.91	10

*Note: The number of observations varies for each aspect of the framework for several reasons: 1) Some teachers were not using all the aspects of the framework in their classrooms; 2) Some aspects of the framework are targeted to early grades (K, 1) and so are not observed in the other grades; 3) On some occasions, observations were interrupted by the school schedule, and the entire literacy block was reduced for that day.

Table 4. Correlations Among Measures for Paired Observations at Time Point 5

Read Aloud	0.91
Shared Reading	0.78
Guided Reading	0.85
Interactive Writing	0.88
Writing Workshop	0.85
Word Study	0.86
Aspects of Teaching	0.80
Teaching for Strategies	0.81

Figure 3. Guided Reading Item Map

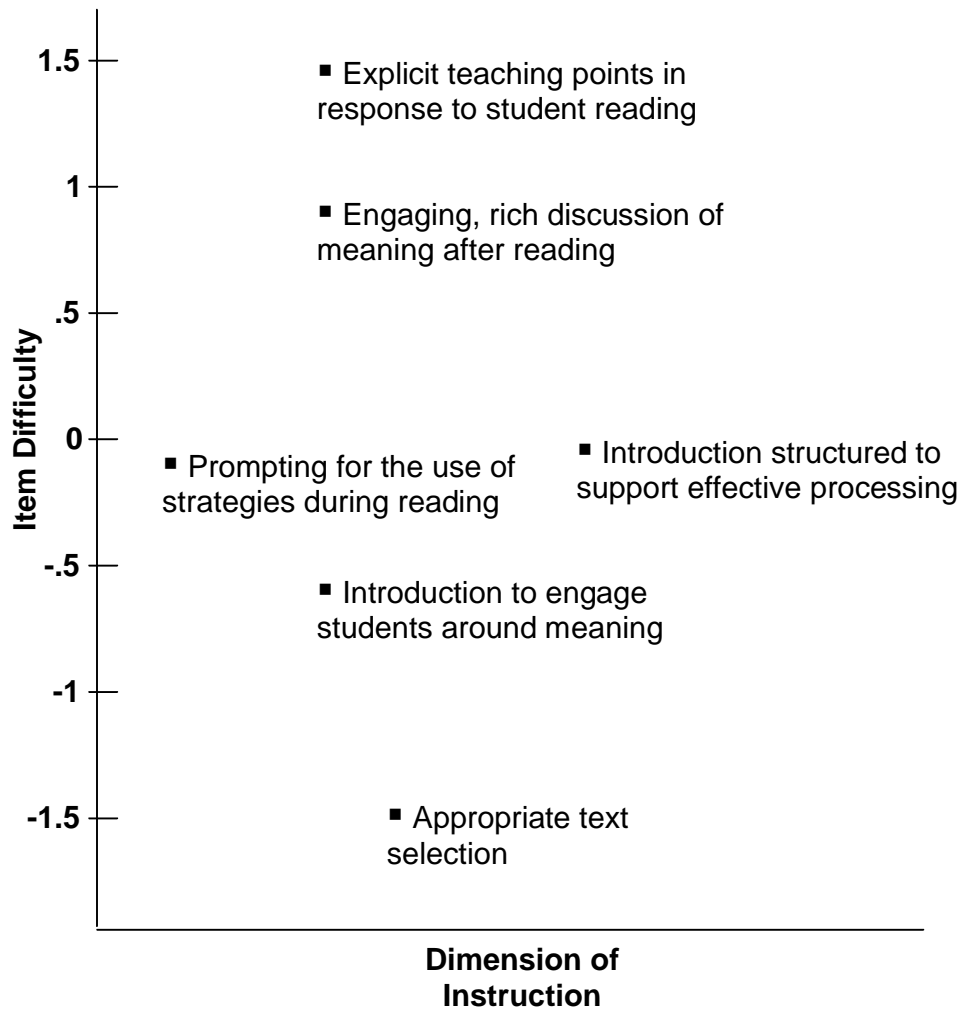


Figure 4. Writing Workshop Item Map

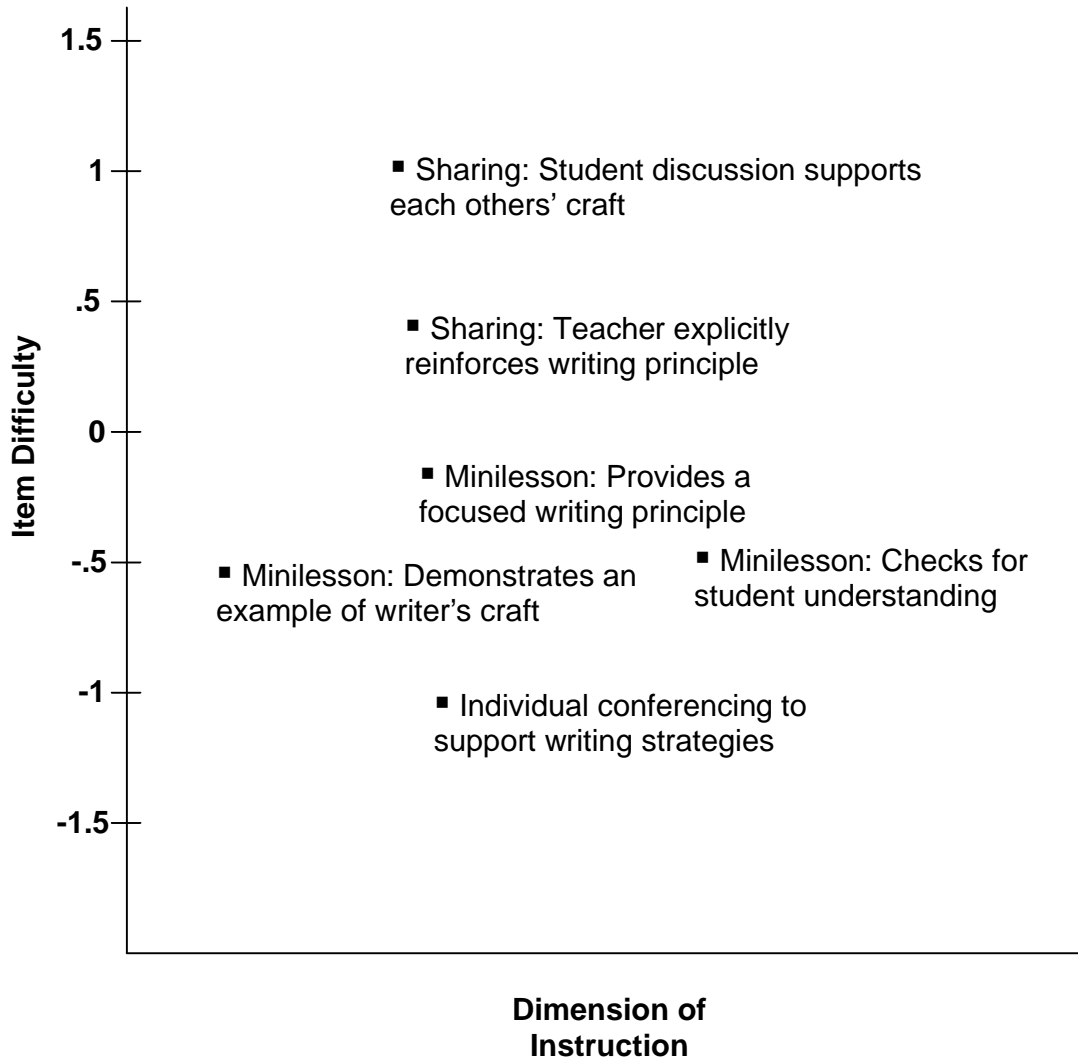


Figure 5. Word Study Item Map

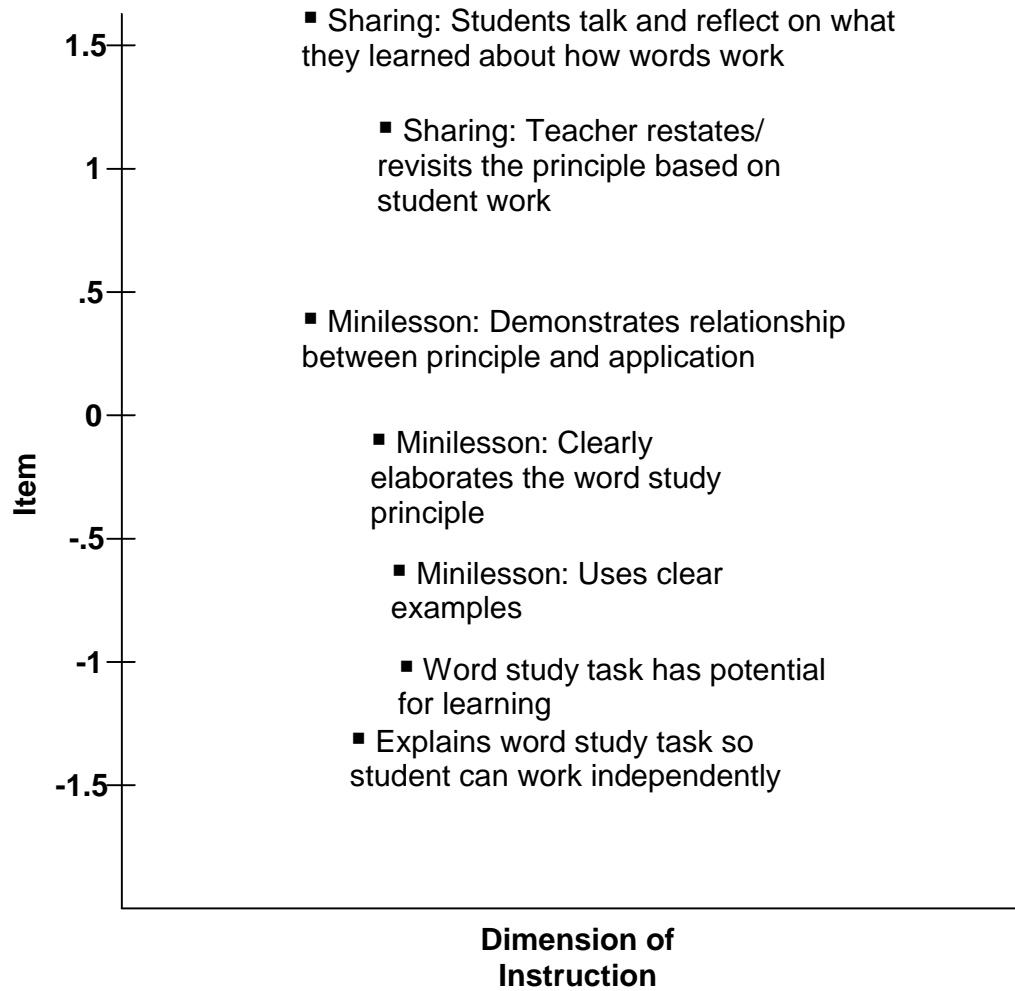
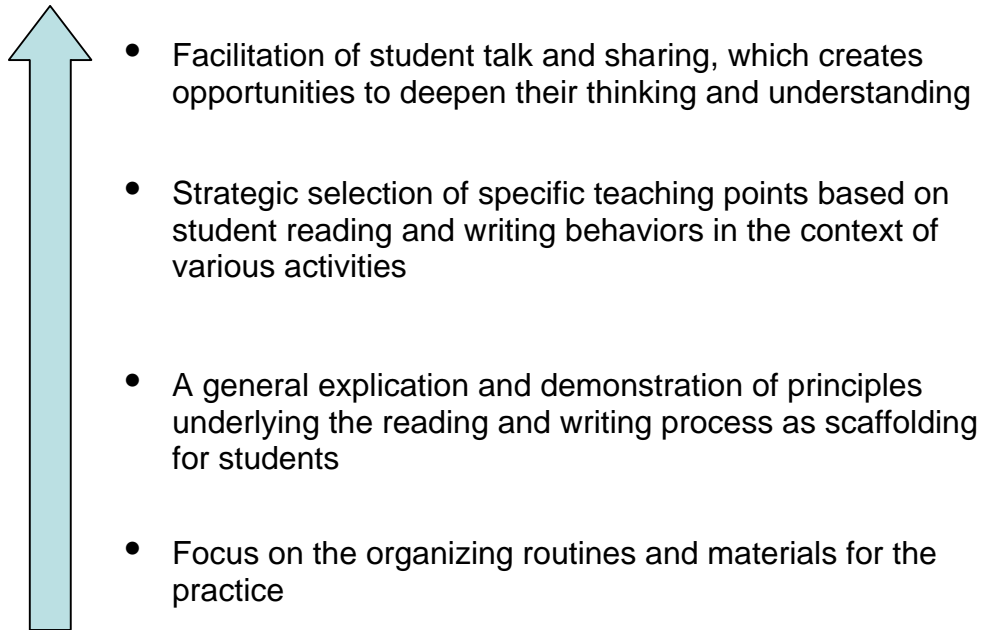


Figure 6. Developmental Model of Teacher Practice



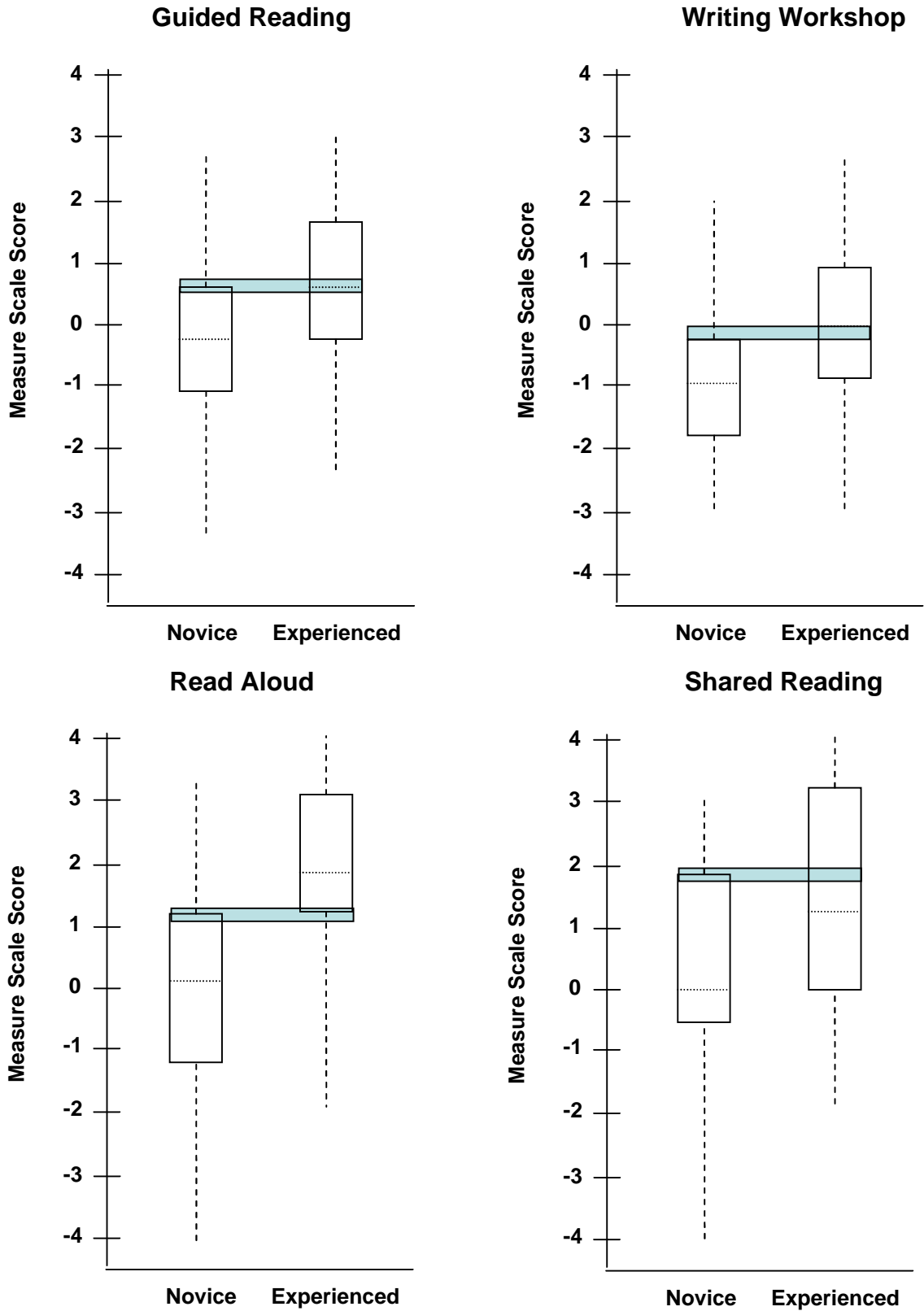
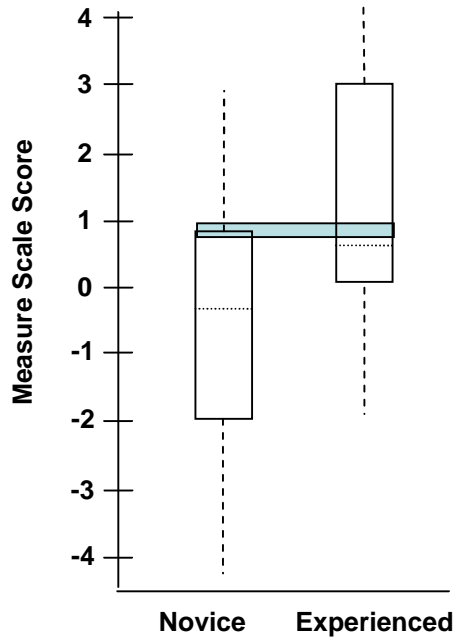
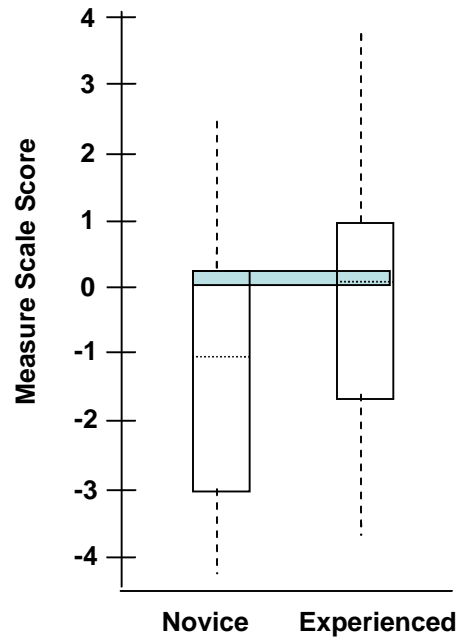


Figure 7. Comparing Novice and Experienced Teachers on the DLLT

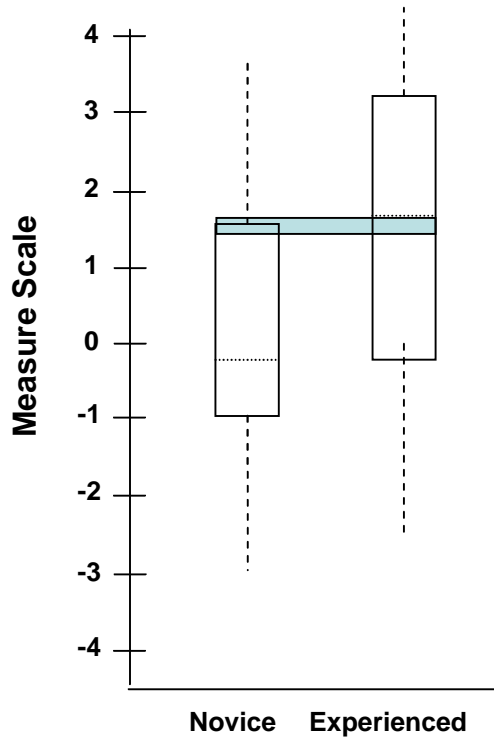
Interactive Writing



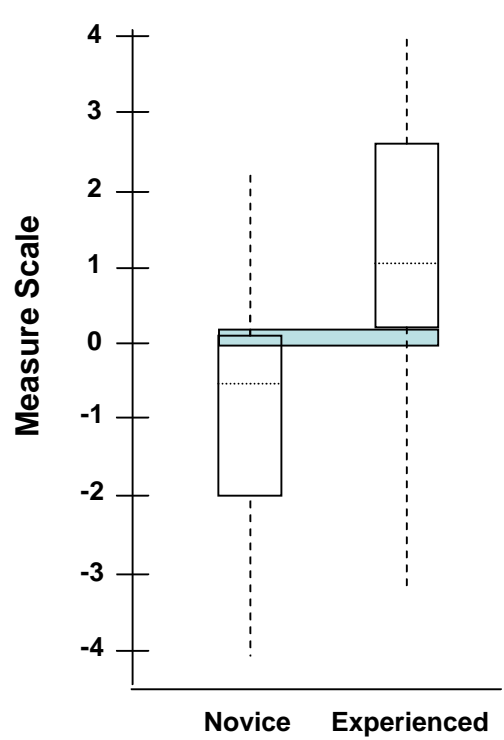
Word Study



Aspects of Teaching



Teaching for Strategies



Appendix A

Part 1—Measurement Statistics

Read Aloud

Person Reliability — 0.63

	Category 4	Item Difficulty	Infit Statistic
RA Row 3	Discussion after reading efficiently builds on overall meaning and extends students' thinking about the text.	0.71	0.95
RA Row 2	Teacher reads aloud and invites interaction; pauses add to the read-aloud session; almost all pauses are very well timed and result in good discussion during reading.	-0.19	0.94
RA Row 1	Teacher engages attention of the students prior to reading with brief comments or questions; prepares students for active listening and response.	-0.52	1.07

Shared Reading

Person Reliability — 0.63

	Category 4	Item Difficulty	Infit Statistic
SR Row 3	Makes appropriate teaching points that extend children's understanding of the reading process. Almost all are clear, specific, and well timed.	1.11	0.98
SR Row 2	Teacher engages almost all children in active shared reading of the text.	-0.33	1.05
SR Row 1	Text is appropriate (language, print, layout, interest) for the age level and the experience of students; text has many learning opportunities.	-0.78	0.92

Guided Reading

Person Reliability — 0.79

	Category 4	Item Difficulty	Infit Statistic
GR Row 6	Makes superbly chosen, specific teaching points that help students engage in effective processing of text.	1.23	1.12
GR Row 5	Engages children in a rich discussion of the meaning of the text that is evident in students' comments about their thinking.	0.83	1.04
GR Row 7	Optional: Shows children something explicit and strategic about how words work. Students are engaged, and there is evidence that they are learning more about word solving.	0.36	0.98
GR Row 4	Samples oral reading and demonstrates, reinforces, and consistently prompts (as needed) for effective reading behaviors and problem solving actions.	-0.08	0.75
GR Row 2	Provides an introduction that includes some or all elements (meaning of whole text, language, aspects of print) in a highly integrated, engaging, and cohesive way.	-0.09	0.86
GR Row 3	Engages students in a conversation that brings them into the text and supports thinking about the meaning of the text.	-0.64	0.99
GR Row 1	Selects a text that is the appropriate level and is very well matched to the group and provides many opportunities to learn.	-1.61	1.16

Interactive Writing

Person Reliability — 0.75

	Category 4	Item Difficulty	Infit Statistic
IW Row 5	Selects a few teaching points that offer new learning without unnecessarily involving children doing what they already know well; children contribute to the writing in ways that have high instructional value.	0.71	0.89
IW Row 3	The teacher engages children in a lively negotiation; options are offered by several children; serious consideration is given to word choice and sequence.	0.34	0.87
IW Row 1	Engages children in interesting experiences and a rich and purposeful discussion before writing;	0.21	0.88
IW Row 4	Keeps the writing moving along at a good pace with superbly selected teaching points; children make contributions that have high instructional value.	0.12	1.02
IW Row 2	Makes writing a highly purposeful and connected activity.	-1.38	1.27

Writing Workshop

Person Reliability — 0.80

	Category 4	Item Difficulty	Infit Statistic
WW Row 5	There is consistent evidence of note taking and continuity from previous conferences.	2.12	2.07
WW Row 6	Provides time for students to share their writing; students comment specifically about other students' writing and show understanding of strategies or craft of writing.	0.54	1.07
WW Row 7	Makes explicit and helpful comments about writing shared by students and clearly reinforces the principle and strategies for writing.	-0.05	1.03
WW Row 1	Provides a minilesson that is clearly stated and focused on a writing principle.	-0.50	0.77
WW Row 2	Provides a clear and explicit demonstration or example of what students need to learn as writers (craft, conventions, or process).	-0.58	0.73
WW Row 3	Checks on understanding of principle or application and elicits comments from students that are evidence of understanding.	-0.58	0.83
WW Row 4	Teacher consistently confers with students; interactions prompt for skillful use of strategies or development of writing craft. Most conferences are focused on helping students learn about the <i>writing process</i> .	-0.89	0.97

Word Study

Person Reliability — 0.83

	Category 4	Item Difficulty	Infit Statistic
WS Row 7	Students actively participate in sharing, comment on their work, and show evidence of learning the principle.	1.83	1.20
WS Row 6	Teacher clearly restates the principle and reinforces learning, through examples of students' work.	1.52	1.15
WS Row 4	Provides an application task that is appropriate and has strong potential for helping students develop greater understanding of the principle.	0.79	0.93
WS Row 3	Clearly demonstrates and explains the application task and explicitly relates it to the principle.	-0.06	0.65
WS Row 1	Provides a minilesson with a clearly and explicitly stated principle <i>or</i> asks children to derive the principle from examples and to state the principle clearly and explicitly.	-0.38	0.86
WS Row 2	Uses good examples; teacher checks for understanding and helps students understand how the principle is related to reading and writing.	-0.98	0.70
WS Row 5	Explains the application task in a way that enables almost all students to perform the task independently.	-1.15	1.12

General Aspects of Teaching

Person Reliability — 0.90

	Category 4	Item Difficulty	Infit Statistic
GAT Row 8	<i>Quality of Interactions:</i> Student discussion builds on the comments of other students; students provide evidence to support their ideas based on the text.	1.50	0.94
GAT Row 4	<i>Student Engagement:</i> Most students are on task almost all of the time; there is a very high level of engagement and purposeful activity.	0.42	0.78
GAT Row 5	<i>Student Engagement:</i> Transitions are orderly and efficient.	0.39	1.00
GAT Row 7	<i>Quality of Interactions:</i> Students have many opportunities to talk to, and learn from, each other.	0.03	1.17
GAT Row 9	<i>Sense of Community:</i> The teacher helps students to take high degree of responsibility for their own behavior and learning and to show respect for the learning of others. (E.g., students know routines and why they use them; they help and treat others with respect.)	0.01	0.87
GAT Row 3	<i>Classroom Materials and Organization:</i> Student/teacher generated charts are accessible, relevant and routinely used by teacher and students to guide learning.	-0.02	1.20
GAT Row 2	<i>Classroom Materials and Organization:</i> Organization works for maximum student independence; use and placement of materials in the classroom is obvious.	-0.59	0.85
GAT Row 1	<i>Classroom Materials and Organization:</i> Materials are highly organized for efficient use by the teacher and students.	-0.71	1.00
GAT Row 6	<i>Quality of Interactions:</i> The teacher consistently listens and responds to students.	-1.03	0.98

Teaching for Strategies

Person Reliability — 0.91

	Category 4	Item Difficulty	Infit Statistic
TS Row 9	<i>Teaching for Fluency and Phrasing:</i> Across reading instruction, teacher demonstrates, attends to, reinforces, and prompts for reading that is fluent, phrased, and well stressed.	0.96	1.30
TS Row 5	<i>Teaching for Inferential and Analytic Thinking:</i> Teacher models his/her own inference and analysis about texts and supports students in using these strategies; explicitly demonstrates how readers can apply these strategies.	0.79	.93
TS Row 6	<i>Teaching for Word Solving:</i> The teacher consistently helps students learn and apply a wide range of flexible and highly effective word solving strategies—recognize words, use word parts, derive their meaning from context.	0.40	0.97
TS Row 3	<i>Teaching for Inferential and Analytic Thinking:</i> Teacher consistently asks questions that extend the meaning of the text and often bring out multiple perspectives; consistently prompts student for evidence from the text that elaborates and supports their answers.	0.21	0.80
TS Row 7	<i>Teaching for Word Solving:</i> The teacher supports students in learning and expanding their understanding of word meanings in multiple contexts. Words are talked about and revisited often.	0.15	0.86
TS Row 10	<i>Teaching for Fluency and Phrasing:</i> Teacher assists children when there is evidence of dysfluent reading in various contexts; teacher avoids interrupting fluent reading.	0.03	1.51
TS Row 8	<i>Teaching for Word Solving:</i> The teacher actively provides instruction on phonemic awareness and/or letter-sound relationships and students have ample opportunity to practice and apply these skills in multiple contexts.	-0.20	1.21
TS Row 2	<i>Teaching for Literal Thinking:</i> Teacher helps students learn how to search for and use information that is in the text.	-0.38	0.91
TS Row 4	<i>Teaching for Inferential and Analytic Thinking:</i> Teacher helps students access and use relevant prior knowledge to understand meaning beyond the literal text; teacher helps students synthesize new knowledge in support of understanding the text.	-0.83	0.88
TS Row 1	<i>Teaching for Literal Thinking:</i> The teacher helps students notice specific information contained in both fiction and factual texts that is vital to the literal understanding of the text and helps them to have an overall understanding.	-1.12	0.76