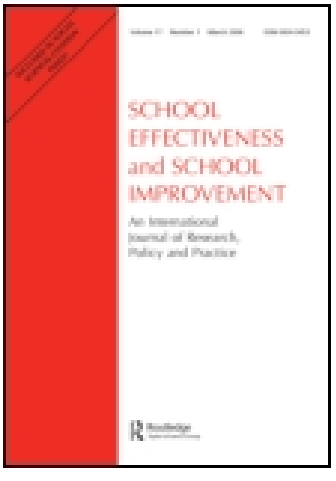

School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nses20

Assessing teacher practice and development: the case of comprehensive literacy instruction

Heather J. Hough [a], David Kerbow [b], Anthony Bryk [c], Gay Su Pinnell [d], Emily Rodgers [d], Emily Dexter [e], Carrie Hung [d], Patricia L. Scharer [d] & Irene Fountas [e]

[a] Public Policy Institute of California, San Francisco, CA, USA
[b] University of Chicago, Chicago, IL, USA
[c] Carnegie Foundation for the Advancement of Teaching, Palo Alto, CA, USA
[d] The Ohio State University, Columbus, OH, USA
[e] Lesley University, Cambridge, MA, USA

Published online: 26 Oct 2012.

To cite this article: Heather J. Hough, David Kerbow, Anthony Bryk, Gay Su Pinnell, Emily Rodgers, Emily Dexter, Carrie Hung, Patricia L. Scharer & Irene Fountas (2013) Assessing teacher practice and development: the case of comprehensive literacy instruction, School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 24:4, 452-485, DOI: 10.1080/09243453.2012.731004

To link to this article: http://dx.doi.org/10.1080/09243453.2012.731004

School Effectiveness and School Improvement, 2013
Vol. 24, No. 4, 452–485, http://dx.doi.org/10.1080/09243453.2012.731004

Assessing teacher practice and development: the case of comprehensive literacy instruction

Heather J. Hough[a]*, David Kerbow[b], Anthony Bryk[c], Gay Su Pinnell[d], Emily Rodgers[d], Emily Dexter[e], Carrie Hung[d], Patricia L.Scharer[d] and Irene Fountas[e]

[a]*Public Policy Institute of California, San Francisco, CA, USA;* [b]*University of Chicago, Chicago, IL, USA;* [c]*Carnegie Foundation for the Advancement of Teaching, Palo Alto, CA, USA;* [d]*The Ohio State University, Columbus, OH, USA;* [e]*Lesley University, Cambridge, MA, USA*

(*Received 14 January 2011; final version received 20 April 2012*)

In this paper, we report on 2 studies developing, testing, and using an observation tool for measuring primary literacy instruction, the Developing Language and Literacy Teaching (DLLT) rubrics. In Study 1 (an instrumentation study), we show that the DLLT has a high level of internal consistency, that there are high levels of inter-rater reliability when the tool is used by in-school coaches, that the items order consistent with a hypothesized developmental trajectory, and that the DLLT can distinguish between novice and more experienced teachers at one time point. In Study 2 (a 3-year longitudinal study), we show that the DLLT is sensitive to growth in teaching practice, that this growth is related to exposure to one-on-one coaching, and that teacher practice as measured by the DLLT is related to teachers' value added to student achievement year by year.

Keywords: measurement; teaching quality; observation rubric

Introduction

There is growing evidence that teacher quality matters (Rivkin, Hanushek, & Kain, 2005; Sanders & Rivers, 1996) and that what teachers actually do in the classroom affects student achievement (D.L. Ball & Rowan, 2004; Grossman et al., 2010; Kane, Taylor, Tyler, & Wooten, 2010). However, even as the field comes to consensus on this fact, there is not yet consensus around what practices contribute most to student learning. Thus, being able to identify effective practices through observation (and specifically to identify those classroom practices associated with high student achievement gains) is central for researchers, policymakers, and school and district leaders looking to increase the effectiveness of the teacher workforce.

It is important to measure instructional quality for a variety of reasons. As Matsumura, Garnier, and Slater (2008) note: (a) school districts making investments in professional development initiatives need methods for evaluating the effects of those initiatives on classroom instruction; (b) including measures of instructional practice to measure teaching quality – as opposed to using just student achievement – may

*Corresponding author. Email: hough@ppic.org

moderate the trend of narrowing the curriculum to only what is tested; and (c) measuring instructional quality encourages educators to engage in more informed discussions about improving instruction. In addition, rigorous evaluations of teaching skill can inform professional development efforts. Teacher organizations, such as the American Federation of Teachers, hold that "a model of continuous professional development based on the growth of individual teachers is the basis of a comprehensive teacher evaluation system" (American Federation of Teachers, 2010, p. 38).

This paper reports on two studies testing the development and use of a new tool for observing literacy instruction, the Developing Language and Literacy Teaching (DLLT) rubrics. The DLLT focuses on primary literacy in grades K-2, as this period is a critical time for beginning readers (Armbruster, Lehr, & Osborn, 2001). At this age, children are taught the skills that together enable them to understand and find meaning in what they read and take advantage of the learning opportunities in upper elementary grades and beyond (Armbruster et al., 2001). Literacy learning in these grades has been linked to later educational achievement, high school graduation, and college attainment. The DLLT aims to measure early literacy instruction taught through a comprehensive framework that emphasizes all of the components of reading, as well as writing and oral language development.

An important feature of the DLLT is that it was designed as a clinical tool to help in-school instructional specialists and leaders, such as literacy coaches, assess teaching and provide individualized support for its improvement over time. We sought to build a reliable and valid tool that could provide coaches with information that would be useful in their work with teachers. To assess the utility of the DLLT rubrics for this purpose, we examined four questions:

(1) Can the DLLT be used reliably by in-school coaches to assess teacher practice in a comprehensive literacy classroom?
(2) Does the DLLT distinguish novice from more experienced practitioners?
(3) Can the tool detect change over time in teacher practice?
(4) Does teacher skill as measured by the DLLT predict student achievement outcomes?

We describe below the conceptual frameworks guiding instrument development and briefly review other extant observational tools. We then overview the development process for the DLLT, and detail two studies that we undertook to examine instrument reliability and validity. In Study 1, we examined the reliably of the DLLT when used by in-school coaches, analyzed whether the DLLT ratings were theoretically interpretable, and assessed whether the DLLT could distinguish between novice and more experienced practitioners. In Study 2, we analyzed data from a large-scale longitudinal investigation in which literacy coaches used the DLLT over a 3-year period. This study allowed us to examine whether the DLLT was sensitive to changes in teaching practices over time; whether these changes were related to coaching exposure; and whether teaching quality, as measured by the DLLT, predicted student literacy outcomes.

Conceptual frameworks

Our goal in developing the DLLT was to create a clinical tool that met psychometric standards. We wanted the DLLT to distinguish reliably among teachers at any given

point in time and to measure changes in teaching practice over time. We conceptualized the latter within the context of expertise development theory. Consistent with a central feature of this theory, that expertise is both content and context specific, the DLLT focuses on measuring teaching in the context of comprehensive literacy in a K-2 setting. We review below the instructional context of comprehensive literacy and then present the model of expertise development undergirding the DLLT.

Context of early literacy instruction

The DLLT is designed to evaluate comprehensive literacy instruction in the primary grades. Fountas and Pinnell (2006) detail this underlying model of literacy processing and literacy acquisition. This theory draws primarily from three areas of theory and research: (a) Clay's (2001) theory and research on the cognitive strategies that students need to acquire to become proficient readers and writers, (b) information-processing theories of how people process complex information (Rumelhart, 1994), and (c) theories of teaching as a process of scaffolding cognitive development through teacher–student interactions and classroom discourse (Cazden, 2001; Tharp & Gallimore, 1991).

Clay (1979, 2001) was one of the first to propose that reading is a process of active problem-solving rather than of passively receiving information from text. Problem-solving activities and behaviors required for reading include searching for information, self-monitoring and self-correcting, making and confirming predictions, and integrating visual, phonological, syntactic, and semantic information in order to understand text. As children gain more skill and automaticity, these actions become integrated into a processing system.

Behaviors involving alphabetic (or other grapheme) knowledge receive explicit attention in Rumelhart (1994). He proposed an interactive theory of reading in which readers must integrate "bottom-up" information from the printed page with "top-down" information stored in memory. Included in this is the process of integrating information about letters with stored phonological knowledge in order to decode written words into spoken words (Goswami, 1999). Readers, however, must go beyond decoding print into spoken words. In order to comprehend text, readers must construct meaning by activating prior knowledge, connecting new information in the text with prior knowledge, and revising their knowledge. The complexity of literacy development is acknowledged in the new Common Core Standards for English Language Arts, which specify grade-level goals for phonics and fluency as well as for literal and inferential comprehension skills (National Governors Association Center and Council of Chief State School Officers, 2010).

Theories of teaching-as-scaffolding suggest that the role of the teacher is to provide support for the student's acquisition of the complex, integrative cognitive strategies required for reading and writing (Fountas & Pinnell, 2006). Some children will acquire these strategies quickly, almost on their own, and others will require more deliberate scaffolding. Instruction-as-scaffolding includes organizing the environment, such as by choosing appropriate books, materials, or activities for the child, and then interacting with the child before, during, and after reading in ways that give the child the opportunity to acquire and practice new strategies. Teachers scaffold student learning by questioning and prompting to help the child access information required for text processing; by modelling and thinking aloud in order to provide the child with a demonstration of proficient processing; and

through explicit instruction in areas such as phonics, which alerts students to patterns in written language and the relationship between letters and sounds.

These forming ideas are advanced through a number of instructional formats commonly found in comprehensive literacy classrooms, including small-group guided reading lessons, reading aloud to children, opportunities to write, and explicit instruction in spelling, decoding, and composing skills (Morrow, 2008). The DLLT was designed to assess each of these specific instructional contexts and also to evaluate overall classroom organization and orchestration of activity to advance student learning. DLLT items consider how well teachers provide the scaffolding necessary for students to acquire the cognitive strategies required for reading and writing. Some items focus on choices teachers make about material, books, and reading and writing activities; others rate how well teachers engage students in discussions that model or prompt strategic thinking; and still other items evaluate teachers' use of direct instruction to explain rules, patterns, or processes.

Teacher development of expertise

The development of the DLLT was informed by core principles about the nature of expertise development from studies in diverse fields. Bransford, Brown, and Cocking (1999) describe several characteristics that distinguish experts from novices, including noticing "meaningful patterns of information" (p. 31), flexibly retrieving knowledge with little effort, possessing deep content knowledge about the subject, organizing knowledge hierarchically into "big ideas" rather than fragmented facts, and adapting flexibly to new situations. They also emphasize that expertise is content and domain specific: Experts may be highly skilled and knowledgeable in some domains but not others.

Flyvbjerg (2001) elaborates five developmental levels in the learning process from novice to expert. The first three levels reflect rule-based thinking and logically based action, or procedural knowledge. At this stage of expertise development, the focus is on following rules and using trial-and-error experimentation, but there is not yet skill in addressing more novel problems or contexts. In contrast, learners at the higher levels of expertise development are defined as "proficient performers who combine intuition and judgment as they identify problems, goals, and plans based in a perspective informed by experience" (pp. 20–21).

In the domain of comprehensive literacy instruction, expert teachers are able to integrate their assessments of their students' literacy abilities, their theoretical knowledge about reading and writing processes, and their knowledge of effective instructional strategies in order to make the moment-to-moment decisions that build toward long-term, high quality instruction (Fountas & Pinnell, 2006). This kind of pedagogical skill entails significant procedural knowledge about particular instructional techniques but also moves well beyond it. In contrast, more novice literacy teachers focus on the procedures and steps of teaching without being able to adjust their teaching to flexibly meet the needs of particular students or particular instructional moments (Neufeld & Roper, 2003).

Other early literacy observation rubrics

A number of classroom observation systems have been created to assess instruction, but most have been designed to measure a teacher's skill at a single time point rather

than how much a teacher *improves or develops* over time. Some rubrics rate teacher behaviors in terms of the *amount* of desired practice that is observed. For example, one section of the Early Language and Literacy Classroom Observation tool (ELLCO) (Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002) measures the number of books read, the number of writing activities, and the number of minutes spent in book reading observed during a 30–40-min period. Juel and Minden-Cupp (2000) developed a system in which observational notes were converted into percentages, such as the percentage of literacy activities that are phonics focused. Other literacy-focused observational systems that are frequency based include Edmunds and Briggs (2003), Estrada (2004), Foorman and Schnatschneider (2003), and Greenwood, Abbott, and Tapia (2003). In these instruments, teacher development over time is conceptualized primarily as an increase in the *quantity* of effective behaviors, rather than as a qualitative change in teaching practices (C. Ball & Gettinger, 2009; Grace et al., 2008).

Other rubrics use Likert scales to rate *how well* teachers perform particular behaviors. For example, the rubrics developed by Graves, Gersten, and Haager (2004) assess instructional effectiveness on a 1–4 scale from *not effective* to *very effective*. Some sections of the ELLCO rate teacher behaviors on a 1–5 scale from *deficient* to *exemplary*. Other such rubrics include the Activity Settings Observation System (Estrada, 2004), the CIERA Classroom Observation Scheme (Taylor, Pearson, Peterson, & Rodriguez, 2005), and the literacy instructional implementation scales developed by Bitter, O'Day, Gubbins, and Socias (2009). Still other rubrics have been developed to evaluate specific individual teaching practices such as scaffolding behaviors during preschool read alouds (Pentimonti & Justice, 2010). In these evaluative schema, teacher development is conceptualized primarily as "getting better" at particular practices. Precise descriptions of hypothesized development trajectories, however, are not explicated.

A few observation rubrics have developed sequenced descriptions of teacher behaviors along a continuum, with qualitative accounts of exemplary teacher behaviors as well as of less exemplary, presumably predecessor, behaviors. These rubrics typically focus on general standards for good teaching as might be seen in any subject and grade. The Standards Performance Continuum (SPC) (Doherty, Hilberg, Epaloose, & Tharp, 2002) rates teachers' behavior related to five standards of effective pedagogy, including one standard related to language and literacy. The SPC has a 5-point scale that assesses teacher behaviors in developmental terms: *not observed*, *emerging*, *developing*, *enacting*, and *integrating*. Teacher behaviors that are "emerging" are those that show that "elements of the standard are implemented," while behaviors that are "integrated" are those showing that "at least three standards are implemented simultaneously" (p. 81). Similarly, the Danielson Framework for Teaching (2007) and the Instructional Quality Assessment (IQA) instrument (Matsumura et al., 2006; Wolf, Crosson, & Resnick, 2006) rate teachers on a continuum of behaviors.

The DLLT differs from the rubrics described above in that it is based on a theory of teacher development from novice to expert applied to the contexts of comprehensive literacy instruction in kindergarten through Grade 2. It draws on theories and research on reading development and is organized around whole class, small group, and individual activities typically associated with comprehensive literacy instruction.

The Developing Language and Literacy Teaching (DLLT) rubrics

In the development of the DLLT rubric system, we sought to use the conceptual frameworks introduced above to design a clinical tool for use in evaluating the quality of teachers' instructional strategies. Because we wanted the tool to be easy to use by practitioners, the item content and specific language were developed in collaboration with expert literacy educators who have been developing and observing coaches for over 15 years. Then, the instrument was piloted, revised, and refined into its current form. This is discussed further in Study 1 below.

Following on the idea that expertise is content and domain specific, we developed separate rubrics for each of the six major instructional activities that form comprehensive literacy. These activities, which involve a strategic matching of learning contexts with instructional goals, are detailed in several major professional resource books used widely by teachers (Beck & McKeown, 2001; Fountas & Pinnell, 1996, 2006; Horn & Giacobbe, 2007; McCarrier, Pinnell, & Fountas, 2000; Pinnell & Fountas, 1998) and include:

(1) **Guided reading** in which the teacher provides instruction to a small, homogeneous group of students;
(2) **Writing workshop** in which students write independently, primarily on topics of their own choosing, and confer with the teacher;
(3) **Word study** lessons in which the teacher provides short, explicit lessons to the whole class on topics such as phonics, spelling, or vocabulary;
(4) **Interactive read aloud** in which the teacher reads a book aloud to the whole class, models fluent reading, and engages the children in discussion at key points in the book;
(5) **Shared reading**, in which the whole class or a small group reads a large-format text aloud with teacher support;
(6) **Interactive writing** in which the whole class or a small group writes a single large-format text with the support of the teacher.

In addition to a separate rubric for each of the activities above, The DLLT also includes two holistic rubrics – one focused on the organization of the classroom and management of basic routines (**General Aspects of Teaching**), and a second on **Teaching for Strategies**. The General Aspects of Teaching rubric is used to assess classroom materials and organization, student engagement, the quality of teacher–student interactions, and sense of community among students across the entire lesson. The Teaching for Strategies rubric was designed to further differentiate truly expert practice from more novice efforts. Within comprehensive literacy, it is argued that 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. Effective instruction demands facility in recognizing and being able to act on the complex interconnections, opportunities, and constraints that may exist among multiple goals being pursued for students, and the instructional activities at a teacher's command to use (Fountas & Pinnell, 2006). Thus, the Teaching for Strategies rubric sought to measure a teacher's capacity to integrate students' learning experiences across the various instructional components that comprise the overall literacy instructional block.

Underlying all eight rubrics is a model of expertise development proceeding from novice or procedural knowledge to more expert enactment. Each rubric consists of a

set of rating items (ranging from 3 to 10) on 4-point scales that offer explicit descriptors of the range of activities likely to be seen in classrooms where teachers are operating at varying levels of expertise in that particular instructional practice.

Overall, 51 separate elements of comprehensive literacy instruction are rated across the 8 rubrics. (See Appendix 1 for an example of the full rubric for Writing Workshop.)[1] Each rating item focuses on a specific, discrete aspect of some instructional activity. The rating categories for each item progress from the absence or basic enactment of a practice, scored at Level 1, to competent practice (scores of 2 and 3), to a description of expert practice at Level 4. See Figure 1 for a depiction of this progression.

The items were developed initially by an expert panel, and then results from a Rasch Rating Scale analysis were reviewed (see discussion below on use of the Rasch measurement model). Items that misfit the measurement model were either deleted or revised, and category descriptors that appeared ambiguous were also revised. Regardless, some of the descriptors, especially for expert (Level 4) ratings, still require a degree of professional interpretation on the part of the rater. As we will show below, we were able to achieve good inter-rater reliability as all raters shared a basic body of professional knowledge about comprehensive literacy instruction and were trained on the use of DDLT.

Each item on the DLLT provides a window into a developmental pathway from novice toward more expert practice. The low end of item descriptors focuses on whether appropriate classroom structures are in place, and the basic instructional routines of comprehensive literacy are being executed. At the next level, item descriptors focus on whether teachers embed in these routines some scaffolding for students about general aspects of the reading and writing processes (as contrasted with just executing the routines). Moving up another level, the items describe instruction where teachers strategically infuse the basic structures and routines of

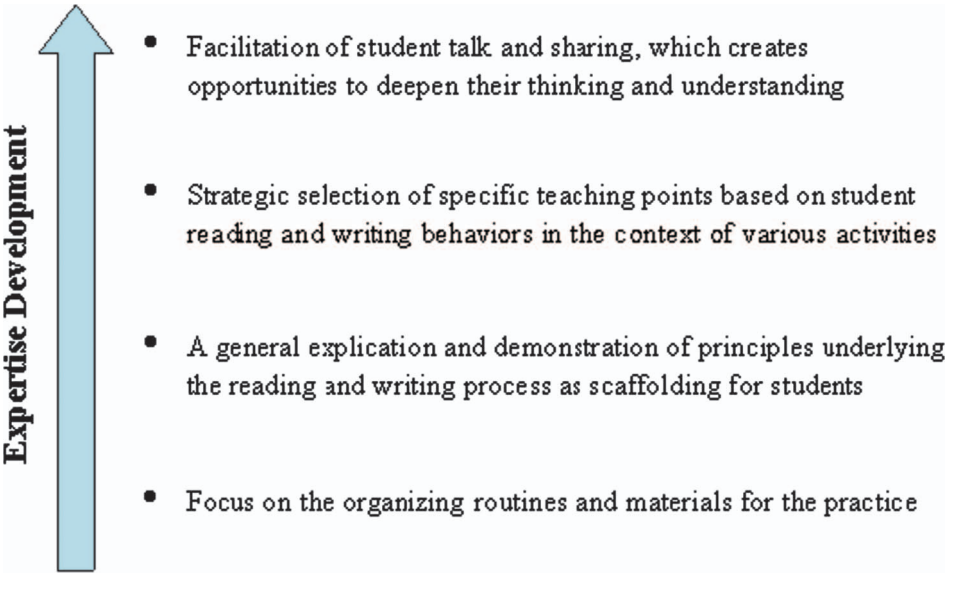

Figure 1. Developmental model of teacher practice.

comprehensive literacy with carefully selected teaching points responsive to observed student behaviors and that are grounded in research-based understandings of reading and writing processes. Finally at Level 4, teachers facilitate student talk about these teaching points to consolidate ''lessons learned.'' Taken overall, the rubrics are structured such that it is generally not possible to achieve a high rating on an item if the lower level descriptors have not also been observed.

The overall goal of the rubric system is to distinguish procedural teaching from more expert practice. Because the DLLT measures 51 separate instructional practices, a teacher may be more or less proficient on different items that comprise each rubric. However, used as an observation system, the DLLT provides an overall literacy profile for any observed lesson. Each rubric can be used separately (e.g., to measure a teachers' Guided Reading practice), or can be used in summary form, to describe a teachers' overall success with comprehensive literacy instruction. Taken together as a set of eight rubrics, the DLLT allows coaches and other support providers to make assessments of specific practices and overall expertise in enacting a complete instructional framework.

Study 1: pilot testing, measure creation, and initial validation

This study was a first examination of the reliability and validity of the rubrics. Specifically, we sought to examine whether the items cohered around the construct of comprehensive literacy, whether the tool could be reliably used by in-school coaches, whether item responses were consistent with a theory of expertise development, and whether the DLLT could distinguish between novice and more experienced teachers in comprehensive literacy. It is important to note that by ''novice'', we refer to teachers who are not necessarily new to teaching, but new to this type of literacy instruction. This initial test of the DLLT was carried out by 22 coaches in 17 schools in 4 districts using a comprehensive literacy instructional framework. In each school, an experienced literacy coach agreed to use the DLLT to observe and document literacy activities in selected classrooms. Training on the instrument involved an introduction to the overall rubric that included discussion of why particular aspects were highlighted for observation. This was followed by observation and individual scoring of a set of video lessons and then group discussion.

Nine of the schools in Study 1 were drawn from a district where 50–89% of the students qualified for Free or Reduced Price Lunch (FRPL). Six schools in two other districts had 47–74% FRPL. Approximately 40% of the students in each of these 15 schools were minority. In contrast, less than 20% of the students in the last 2 schools in the fourth district qualified for FRPL, and about 20% were minority. Note this is a different sample of schools than participated in Study 2 described below.

The design called for each coach to observe four teachers, one at each grade level K-2, on five occasions (once a month) from January through May of 2005. The goal was to explore how well the instrument functioned across the range of teachers that coaches typically engage. Coaches within each school were asked to select two novice teachers (2 or fewer years of experience teaching within a comprehensive literacy framework) and two experienced teachers (3 or more years of teaching within the framework) to include in the study. In each of the five observations, we asked the coaches to observe a full literacy block (between 90 and 150 min) and rate each lesson using the DLLT rubric system. In order to obtain data on inter-rater

reliability, paired observations were designed for both the initial and final time points. On these occasions, a participating coach from one school travelled to another school in the same district so that the two coaches could jointly observe, but independently rate, the same lesson.

In order to minimize the overall data collection burden on the participating literacy coaches, the paired observations conducted at Time points 1 and 5 counted toward coaches' commitment to conduct up to 20 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.

Data

Due to participant attrition, scheduling conflicts, and the limited number of classrooms and target teachers in small schools, the implemented study deviated somewhat from the original design. In the end, we obtained useable rubrics from 275 literacy block observations. These observations came from 78 classroom teachers in the 17 schools. Of this group, 37 teachers were novices within a comprehensive literacy framework, and 41 were more experienced. The average coach observed 3.5 teachers, and the average teacher was observed on 3.5 occasions. In addition, paired observations were conducted on 33 lessons at Time points 1 and 5, respectively. Each coach was involved in at least one joint observation at both of these time points.

Methods

A primary analytic task for this study was to examine the internal consistency and inter-rater reliability of each scale. We also sought to examine whether the items that comprise each of the eight separate rubrics form a theoretically interpretable and empirically defensible scale, as hypothesized by the conceptual frameworks. For this purpose, we analyzed our data separately for each of the eight rubric domains using a Rasch Rating Scale analysis (Wright & Masters, 1982). The analysis produces a "difficult estimate" for each item. Formally, this is the log odds of a teacher scoring in a given item category relative to the base category for that item (a score of 1). Applied to the DLLT rubrics, these item difficulty estimates provide empirical evidence about the relative prevalence of more expert practice among the set of items that comprise each rubric. An item with a high difficulty estimate means that fewer teachers were rated highly (e.g., Categories 3 and 4 versus 1 and 2) on that particular practice as compared to other items in the rubric with lower difficulty estimates. Thus, the Rasch analysis empirically arranges items in a hierarchical order scale and then uses this scale as the basis for assigning a scale score for each teacher in the same log odds metric as formed by the items. The resultant measures characterize each teacher's level of expertise on the underlying construct. In the context of our study, the hypothesized underlying construct is "teaching expertise" within the framework of comprehensive literacy.

The Rasch analysis also produces a separate "infit statistic" for each item, which allows us to examine whether each item actually operates in a way consistent with its location (i.e., difficulty placement) in a hierarchical scale. That is, if the item is functioning properly, a teacher scoring high on an item should also be scoring highly

on the easier items in the scale but have a lower probability of a high score (e.g., Categories 3 and 4) on the more difficult items. Thus, infit statistics provide a basic check on the unidimensionality of each rubric operating as a hierarchical scale.

Results

Reliability

We examined both internal consistency reliability and inter-rater reliability.

Internal consistency reliability. The DLLT was designed to assess instructional practice in the six specific instructional areas and two holistic areas detailed above. As part of a Rasch Rating Scale analysis, we calculated alpha reliability coefficients for each of the eight separate rubrics that compose the DLLT. The alpha coefficients, also referred to as person separation reliabilities (Wright & Stone, 1979), provide evidence as to the internal consistency of the set of items used to assess the targeted construct. They also tell us how well the resultant summary measures, based on these items, reliably differentiate among the individuals observed. Overall, the rubrics demonstrated good internal consistency reliability (see Table 1). The six rubrics that consisted of five or more items achieved internal consistency reliability of 0.75 or higher. For the other two rubrics, both of which consisted of only three items, the internal consistency reliability was 0.63.

Inter-rater reliability. The Rasch rating scale model produces a composite score, called a measure, for each separate rubric by aggregating together the responses to the individual items that compose the rubric (Wright & Masters, 1982). These Rating Scale measures are analogous to the summary scores produced by item response theory models, as used in most standardized testing programs such as the SAT and ACT. To calculate inter-rater reliability, we computed correlations separately for each of eight rubric measures from the paired observations conducted at Time point 5. All of these correlations were 0.80 or higher, except for Shared Reading, which

Table 1. Internal consistency reliability of the DLLT rubrics.

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

was 0.78 (see Table 2). These results indicate quite good inter-rater agreement at the rubric measure level, especially given the high inferential nature of some of the rating scale descriptors.

Validity

For the purposes of our initial test of the validity of the DLLT, we examined whether (a) the rating items ordered within each rubric in ways consistent with the theoretical and practice frameworks used to develop each item set, and (b) whether the tool distinguished between teachers with varying levels of experience in comprehensive literacy instruction.

Item functioning. Underlying the DLLT is a theory of expertise development. If the DLLT rubrics are truly measuring teaching practices along an arc of development, individuals who are rated high on a particular item should be more likely to be rated high on the "easier items" that are below it in the scale and be less likely to demonstrate competence on the "more difficult" to master items that appear higher in the scale. In general, the item-difficulty statistics (see Appendix 2) are consistent with the hypothesized model presented in Figure 1. Within each rubric, items with lowest item difficulty tend to focus on the organizing routines and materials for the practice; items with the highest item difficulty are more likely to emphasize the skillful organization of instruction such as facilitating student discussion that creates opportunities for deep thinking and comprehension.

The item infit statistics provide a second source of evidence for examining how well each item set functions as a hypothesized developmental scale. Within each rubric, we would expect infit statistics to range from 0.5–1.5, if these item sets are operating as hierarchically ordered (Linacre, 2010). Indeed, as shown in Appendix 2, only one of the 51 rubric items across the eight scales falls outside this range.

The combination of item difficulty and infit statistics provides evidence that the scales appear to function as a general developmental progression for most teachers. If they did not, more extreme infit statistics would have been observed. These results provide evidence that it is appropriate to view each rubric as a developmental scale from novice to more expert practice; this hypothesis is further tested below.

DLLT distinguishes between novice and experienced practitioners. As a first test of the hypothesis that the DLLT assesses teachers' development of instructional expertise

Table 2. Correlations among measures for paired observations at Time point 5.

Rubric	*Correlation*
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

over time, we examined whether the DLLT distinguished between novice and more experienced teachers in comprehensive literacy instruction. For this purpose, we used the composite measures generated for each separate rubric observation by the Rasch Rating Scale analysis. Basically, the Rasch measures aggregate together the responses to the individual items that compose each rubric while taking into account the relative difficulty of each item category. These rubric composite scores are expressed in the same log-odds metric as the item difficulties that form the underlying scale. We compared measures from teachers relatively new to comprehensive literacy instruction with those from teachers who had more experience. Approximately half of the teachers in Study 1 were relatively new to the framework (1 to 2 years of use) and the other half more experienced (3 or more years). Figure 2 presents a set of box plots comparing these two groups on the eight separate measures. All of the mean differences displayed here are statistically significant beyond the 0.001 level.

In general, we found large differences between these two groups on each of the six core instructional activities that are typically found in comprehensive literacy instruction. For example, in Guided Reading and Writing Workshop, the median score for the group of teachers experienced with comprehensive literacy is equivalent to roughly the 75th percentile in the distribution for the teachers who are

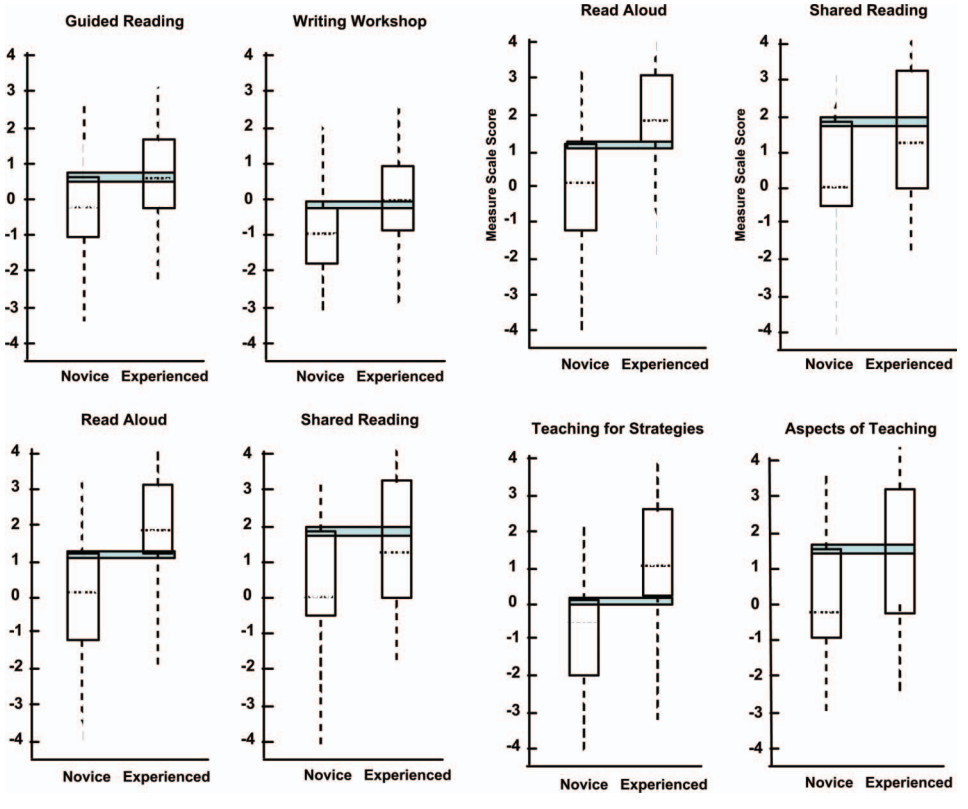

Figure 2. Comparing novice and experienced teachers on the DLLT.
Note: The scale units are measured in logits (i.e., the log odds of being in a particular score category on a given element relative to an overall base category).

novices. Similar size differences exist for the holistic rubric, General Aspects of Teaching.

An even larger difference appeared when we compared novice and experienced teachers on the Teaching for Strategies rubric. On this rubric, the 25th percentile among experienced teachers scored at about the same level as a 75th percentile for novice teachers. 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 evaluate how well teachers integrate their efforts across all elements of the framework. Conceptually, this is a direct measure of expert teaching, and not surprisingly, it discriminates the most between teachers who are novices versus more experienced in comprehensive literacy instruction.

Study 2: assessing teacher development

In Study 2, we sought to determine whether the DLLT could detect development in teacher practice over time, if these changes were related to exposure to literacy coaching and to changes in student achievement. We investigate these questions in the context of a larger study of the effectiveness of the Literacy Collaborative (LC), which is a comprehensive literacy instructional model that builds on the work of Clay (2001) and Fountas and Pinnell (2006). LC is a K-6 school reform project that has been implemented in more than 1,000 elementary and middle schools across the United States. The LC program focuses on training a school-based literacy coach who provides long-term professional development and coaching (for more information about Literacy Collaborative, see www.literacycollaborative.org).

A 4-year study was conducted in 17 schools across eight states in the eastern half of the United States. During the baseline year of the study (2004–2005), one coach in each school was trained in the LC model, but no school-based professional development activity was yet initiated by them. Actual classroom coaching and the collection of DLLT data began in Year 2. Each coach was responsible for all of the literacy coaching of K-2 teachers in her respective school.

In the sample schools, 45% of the students were low income, 16% were African American, 6% Latino/a, and 7% Asian. Schools varied widely in their student composition. In several schools, more than 90% of students were White, while in other schools 30% or more students were African American or Latino. Similarly, the schools ranged in their socioeconomic make-up, with the percentage of students receiving free- or reduced-priced lunch ranging from a low of 19% to a high of 86%. In terms of teacher demographics, participating K-2 faculty varied from as few as 4 teachers in a small primary school to a high of 23. Over 90% of the teachers in most schools were White with on average about 10 years of teaching experience prior to initiation of LC. About half of the teachers (52%) remained in their schools for all 3 years of the study. This varied from a low of 23% in one school to a high of 81% at another.

Data

The larger study design employed a multifaceted data collection approach to explore the linkage between coaching, changes in teacher practice, and changes in student learning. To address each of these causal connections, the project collected

coaching logs documenting teacher exposure to one-on-one coaching, collected multiple waves of the DLLT to measure teacher practice over time, and measured student learning gains (fall to spring) in each classroom for each academic year. Over the 3 years of program implementation (2005–2006 to 2007–2008), 27,427 literacy assessments were collected on 8,576 students taught by 287 teachers. The study also administered surveys asking questions about teacher attitudes, beliefs, and experiences both prior to each teacher's introduction to LC and at the end of the study. These data are not used in this article and are not described here.

Exposure to coaching

Starting in the second semester of Year 1 of LC program implementation, each literacy coach kept a record of the one-on-one coaching sessions she conducted with each individual teacher in the school. These records were monitored month by month by the research team to ensure continuous and reliable recording of each coach's activities. For the analyses reported on in this paper, we created a simple summary of the average number of coaching sessions received per month by each teacher who was present in the school and eligible for participation. As Atteberry and Bryk (2011) have documented, substantial variation was found among teachers in exposure to coaching. On average, teachers received 0.74 one-on-one coaching sessions per eligible month with a standard deviation of 0.45.[2] Thus, while a few teachers received no coaching, others received 1.5 or more sessions per month (or around 13 coaching sessions in a 9-month school year).

Developing teaching expertise: the DLLT

Coaches were asked to use the DLLT to document the instructional practices of teachers in their classrooms. Coaches were trained intensively on the use of the DLLT at the beginning of the 1st year of LC program implementation. This training involved an introduction to the eight rubrics that included discussion of why particular aspects were highlighted for observation. Coaches observed video lessons as a group, scored them individually against the rubrics, and then engaged in group discussions to clarify appropriate rubric use. As a final step, each individual coach rated multiple instructional videos to assure scoring reliability against expert judgement. In order to ensure continued reliability throughout the study, prior to Years 2 and 3, coaches attended supplemental training on the DLLT. In addition, during occasional school-site visits, LC trainers co-observed and rated classes with coaches as further assurance of reliability. LC field staff reported no problems with coaches' use of the DLLT after Year 1.

In order to capture possible changes in teaching practice over time, we asked coaches to complete DLLT ratings for each K-2 teacher in the fall, winter, and spring of each year. Ideally, each teacher would have nine observations (fall, winter and spring for 3 years), but many have far fewer. This is mostly due to normal processes of teacher mobility and attrition. Only 52% of the teachers taught continuously for all 3 years in a K-2 classroom in study schools. In addition, coaches began reporting on each teacher as they initiated LC professional development with that teacher. Especially in the larger study schools, teachers' participation in the program phased in over the course of the first 2 years due to a constraint in how many teachers a coach

was able to initiate work with simultaneously. Teachers who taught for 1 school year have an average of 2.22 DLLT ratings; those who taught for 2 years have 4.89 ratings; and those who were present all 3 years have an average of 7.57 DLLT ratings. Over the 3 years that data were collected, 88% of eligible teachers have at least one rating, with a total of 1,317 observations for 219 teachers.

In Study 1, each of the eight subscales on the DLLT was subject to separate Rasch Rating Scale analyses. The results of these scaling processes specify empirically the arc of development from novice toward greater expertise in each instructional practice. As a complement, Study 2 sought to use these scales to examine teachers' overall development in their instructional practice over time. For this purpose, we formed a composite measure of teacher practice that we call *expertise-in-enactment*. All of the 51 items were scaled together to form this composite index to chart teachers' overall development over time. Again, we examined the item difficulty statistics and infit statistics to assess adequacy of model fit, and no evidence was found of significant violation of model assumptions. The resultant measures were deemed adequate for the purposes of this study. The estimated standard error of measurement associated with these composite scores were incorporated into the analyses through Level 1 measurement models. The specific models used are detailed below.

Student achievement

Approximately 1,150 students were assessed in the fall and the spring of each grade in kindergarten through second grade in each year of the study. This represents a student participation rate of 90% or higher at each testing occasion. We used two reading assessments in order to assess broadly students' literacy learning over the primary grades in this study: the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and the Terra Nova Multiple Assessments of Reading. The DIBELS taps a range of early literacy skills, including letter recognition, phonological awareness, decoding, and oral reading fluency (Good & Kaminski, 2002). The Terra Nova is a group-administered, standardized, norm-referenced reading test (McGraw-Hill, 2001). As reported on in other published research (Biancarosa, Bryk, & Dexter, 2010), these two tests were scaled together using Rasch modelling (Wright & Masters, 1982). The resultant vertical scale provided a basis for examining student literacy learning gains over time, redressing difficulties associated with floor and ceiling effects in using DIBELS assessments in program effects studies.

Methods

This study considered three separate questions: (a) Is there evidence of an overall change in teaching practice and is there variation among teachers in growth rates over time? (b) If variation exists, is it related to exposure to coaching? And (c) Is measured teaching practice related to student learning in gains in these classrooms? To answer each of these questions, we used hierarchical linear modeling (HLM), since our basic data structures involve repeated observational reports on teachers nested within schools. In addition, HLM takes into account the varying amounts of data collected on each person (Raudenbush & Bryk, 2002).

Results

Teacher growth in expertise-in-enactment

We examined the variation in DLLT scores within teachers to assess how much of the observed variability represented actual change in practice over time. In developing the time metric for this growth model, we created a "clock" for each teacher that started when she was first exposed to the LC intervention. For teachers in the school during Year 1 of implementation, their clocks began at the beginning of that year. (A basic literacy workshop occurred during the fall of the first semester, and individual coaching did not begin in most classrooms until the second semester.) For teachers that entered the school in subsequent years, their clock began at the beginning of that year. More specifically, we set TIME = 0 at the data collection time point when a teacher first became eligible for LC coaching; then TIME counts up from there in terms of the number of subsequent scheduled observation occasions. By structuring TIME in this fashion, the intercept in the growth model captures teachers' initial DLLT score as exposure to the instructional framework begins. The slope coefficient represents the rate of change in a teacher's outcome per observation window.

Figure 3 presents the observed mean expertise-in-enactment measures across all nine possible observation occasions; each possible occasion is a quarter within the school year (fall, winter, spring). Data are in the logit metric generated through Rasch measurement models. On average, teachers scored 0.23 on the logit scale during their first observation period, and 0.96 at the end of the study (spring observation in Year 3). This is a change of 0.70 standard deviations.

To model possible changes in teaching practice, we used a 4-level HLM (version 7.0), where unreliability due to scale measurement error is represented at Level 1, repeated measures within teachers are represented at Level 2, teachers are represented at Level 3, and schools are represented at Level 4. We also added a dummy variable (SPR06) at the repeated-measures level, to control for a measurement artefact associated with the 1st-year spring DLLT scores;[3] a dummy variable (SCH25) to control for a school that experienced a change in coach during the course of the study, which introduced additional variability in the DLLT reports from that school; and variables to control for whether the teacher's 1st year in the

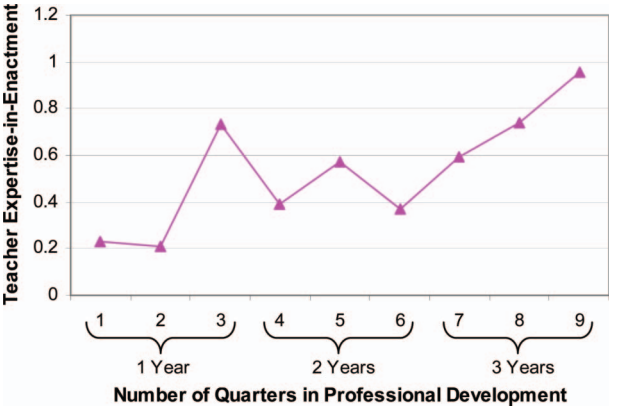

Figure 3. Observed mean growth in expertise-in-enactment.

school was after the 1st year of implementation (Y2ENT and Y3ENT). At both the teacher and school level, the intercept was allowed to vary randomly. Figure 4 shows this baseline model. In this baseline model, there are three sources of parameter variation: observations within teachers, among teachers within the same school, and between schools. Forty-six percent of the total variation in DLLT measures is among time points at Level 2. Of the remaining variation, 38% is among teachers within school and 16% between schools. (See Appendix 3 for full model results.)

In modelling growth, we included TIME as a predictor at Level 2, and allowed the intercept and TIME slope to vary randomly at both the teacher and school level. We found that the inclusion of TIME as a linear predictor in this model explains 63% of the variance at Level 2 (i.e., among the repeated observations). This indicates that there is substantial variation among teachers in the rate of change in their instructional practice over time. The modelled average growth rate (which controls for year of entry and other variables as identified above) is 0.16 with a 95% plausible value range of 0.09 to 0.23 per observation window. The variation in growth rates is statistically significant; the chi-square test statistic for the hypothesis that there are no individual differences among growth rates is 264.19 ($df = 195$, $p = 0.001$). (See Appendix 3 for full model results.)

Teacher growth in expertise-in-enactment related to exposure to coaching

Having established that there is significant growth in teacher development and substantial variability in growth rates among teachers, we investigated next whether

Measurement Model

$$(DLLT/WGT)_{mtij} = \Psi_{1tij}(1/WGT_{mtij}) + \varepsilon_{mtij}$$

Repeated Measures

$$\Psi_{1tij} = \pi_{10ij} + \pi_{11ij}(SPR06) + e_{1tij}$$

Teacher Level

$$\pi_{10ij} = \beta_{100j} + Y2ENT_{ij} + Y3ENT_{ij} + r_{10ij}$$
$$\pi_{11ij} = \beta_{110j}$$

School Level

$$\beta_{100j} = \gamma_{1000} + \gamma_{1001}(SCH25_j) + u_{100j}$$
$$\beta_{101j} = \gamma_{1010}$$
$$\beta_{102j} = \gamma_{1020}$$
$$\beta_{110j} = \gamma_{1100} + \gamma_{1201}(SCH25_j) + u_{110j}$$

Figure 4. Baseline model for observing variation in teacher expertise-in-enactment.
Note: Consistent with the use of Level 1 in an HLM as a measurement model that incorporates into the analysis estimated standard errors of measurement, the Level 1 variance is constrained to be fixed at a value of 1.0. For further discussion, see Raudenbush and Bryk (2002, p. 249).

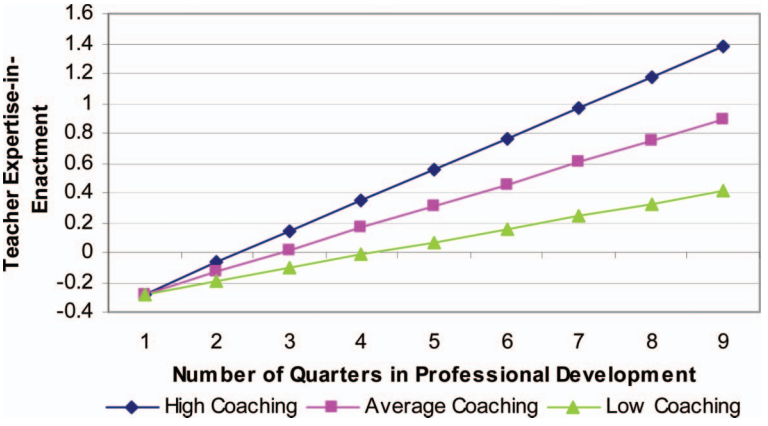

Figure 5. Growth in expertise-in-enactment, by exposure to coaching.
Note: This is based on results from the HLM growth model. The figure illustrates the predicted differences in teachers growth in expertise-in-enactment between those experiencing low amounts of coaching (one standard deviation below the mean) and those exposed to more intensive coaching (i.e., "high" defined as one standard deviation above the mean).

the variability in growth was related to exposure to one-on-one coaching. Others have examined how teacher background characteristics in the LC study are related to the amount of coaching received (Atteberry & Bryk, 2011). Here, we examined the relationship between change in expertise-in-enactment and the amount of exposure to coaching. If the DLLT is actually measuring expertise development and if the latter is related to coaching, we would expect to find a positive association between these two variables.

To test the hypothesis that exposure to coaching is related to growth in teacher practice, we included exposure per month to coaching as a predictor in Level 3 (the teacher level) of our HLM model.[4] We found a statistically significant, positive association (coefficient = 0.13, *se* = 0.05, *p* value = 0.007). A teacher who received extensive coaching (defined as one standard deviation above the average (equivalent to an additional 0.45 coaching session per month) had a growth rate on the DLLT of 3.41 standard deviations higher than a teacher who received minimal coaching (one standard deviation below average). See Figure 5 for a visualization of this effect and Appendix 3 for full model results. For further analysis of variability in exposure to coaching, see Atteberry and Bryk (2011).

We caution that we do not know whether the observed relationship is causal in that the amount of coaching was not randomly assigned to each teacher, and no claim is made to this effect. Rather, these results simply document that the observed association is consistent with what we would expect to see if the DLLT is actually measuring changes in teacher practice over time. As such, they provide one additional small test for assessing construct validity of the DLLT.

Expertise-in-enactment related to student achievement

We have shown that teachers' expertise-in-enactment increases over time and that the variability among teachers in their rates of change is related to exposure to one-on-one coaching. In a final validity test, we examined whether teachers' annual

expertise-in-enactment scores were significantly correlated with estimates of the value-added to student learning in reading in those same classrooms year by year. If the DLLT is measuring the quality of instruction, we would expect larger student learning gains in classrooms with higher teacher ratings.[5]

To estimate the value added by each teacher each year, we build off of the work of Atteberry and Bryk (2011). The authors used a hierarchical crossed-random effects model to assess student literacy learning over 3 years of LC program implementation against observed growth under baseline conditions. They found that, on average, children in participating schools in the 1st year of implementation made 16% larger learning gains than observed during the baseline "no treatment" period. In the 2nd year, children were learning 28% more as compared to baseline data, and by the 3rd year they were learning 33% more. Using the same data and analysis model developed by the authors, we reran their model to create Empirical Bayes estimates of the "value added" to student learning by each teacher each year. The value added is the residual gain among a classroom of students in a given year compared to the baseline, taking into account all prior test results from the classroom's students and other fixed effects included in the model.

We then sought to correlate these Empirical Bayes residuals with the annual expertise-in-enactment rating on the DLLT for each teacher. In order to do so, we first created an average annual measure on the DLLT for each teacher, using a four-level HLM model that allowed us to pool the multiple DLLT measures within each academic year (Level 2), include information on the estimated standard errors for each DLLT measure (Level 1), and time-trend adjustments for any missing DLLT observations (Level 3). This HLM model generated an Empirical Bayes DLLT residual score for each teacher in each year that he/she participated in the study.

Measurement Model

$$(DLLT/WGT)_{mtij} = \Psi_{1tij}(1/WGT_{mtij}) + \varepsilon_{mtij}$$

Repeated Measures

$$\Psi_{1tij} = \pi_{10ij}(\text{Year1}) + \pi_{11ij}(\text{Year2}) + \pi_{12ij}(\text{Year3}) + e_{1tij}$$

Teacher Level

$$\pi_{10ij} = \beta_{100j} + r_{10ij}$$
$$\pi_{11ij} = \beta_{110j} + r_{11ij}$$
$$\pi_{11ij} = \beta_{120j} + r_{12ij}$$

School Level

$$\beta_{100j} = \gamma_{1000} + \gamma_{1001}(SCH25_j) + u_{100j}$$
$$\beta_{110j} = \gamma_{1100} + \gamma_{1101}(SCH25_j) + u_{110j}$$
$$\beta_{120j} = \gamma_{1200} + \gamma_{1201}(SCH25_j) + u_{120j}$$

Figure 6. Model for creating annual expertise-in-enactment scores.
Note: Consistent with the use of Level 1 in an HLM as a measurement model that incorporates into the analysis estimated standard errors of measurement, the Level 1 variance is constrained to be fixed at a value of 1.0. For further discussion, see Raudenbush and Bryk (2002, p. 249).

Measurement Model

$$(SCORE/ERROR)_{mtij} = \Psi_{1ti}(1/VA\_ERROR_{mti}) + \Psi_{1ti}(1/DLLT\_ERROR_{mti}) + \varepsilon_{mti}$$

Repeated Measures

$$\Psi_{1ti} = \pi_{10i} + e_{1ti}$$
$$\Psi_{2ti} = \pi_{20i} + e_{2ti}$$

Teacher Level

$$\pi_{10j} = \beta_{100} + r_{10i}$$
$$\pi_{20i} = \beta_{200} + r_{20i}$$

Figure 7. Model for correlating annual DLLT and value-added scores.
Note: Consistent with the use of Level 1 in an HLM as a measurement model that incorporates into the analysis estimated standard errors of measurement, the Level 1 variance is constrained to be fixed at a value of 1.0. For further discussion, see Raudenbush and Bryk (2002, p. 249).

Level 4 captured the nesting of teachers within schools. Figure 6 displays the model used to create these scores.

Using the Empirical Bayes annual expertise-in-enactment and the value-added scores, and estimated standard errors of measurement associated with each, we than ran a three-level hierarchical linear model that allowed us to take into account at Level 1 the differential unreliability in these respective indicators over classrooms and occasions. Specifically, the outcomes in these analyses are the Empirical Bayes (EB) classroom-level indicators for teachers' expertise and value added to student learning each year. As is customary with use of measurement models at Level 1 in an HLM (Raudenbush & Bryk, 2002, p. 245), the Empirical Bayes estimates are weighted inversely proportional to their respective posterior Bayes standard errors. This results in a Level 2 model with two outcomes: latent values for the DLLT and value-added residuals classroom by classroom and year by year. Each of these was allowed to vary randomly at the year (Level 2) and teacher levels (Level 3) yielding a consistent estimate of the correlation between the expertise and value-added residuals. The second level also included variables indicating the three separate years of the study as fixed effects. (Note that a school level was unnecessary because Empirical Bayes residuals were the deviation for a particular classroom in a particular year compared to the average for the school in that year.) See Figure 7 for the model used.

Results indicated that the classroom expertise and value-added effects were correlated at 0.33 when pooled across teachers and the 3 years of the study. (To place this result in perspective, the Measuring Effective Teaching (MET) Project (Bill and Melinda Gates Foundation, 2010) reports correlations across 2 years in teachers' value added to student literacy learning of the same magnitude.) This means that year by year, teachers who had high expertise ratings tended to have higher value added, and those with low expertise scores tended to have lower value added. We conclude that at any point in time for any given teacher in any given school, expertise in enacting the LC model is moderately associated with that teacher's value added to student learning that year. Finding a statistically significant association here provides further evidence supporting the validity of the DLLT, especially given the relatively small effect sizes of other studies seeking to understand the relationship

between specific teaching practices and student achievement (see, e.g., Grossman et al., 2010; Tyler, Taylor, Kane, & Wooten, 2010).

Summary, conclusions, and limitations

In our work, we sought to develop a tool for assessing teacher practices in early literacy that can be used reliably by in-school coaches, that can measure growth in teacher practice, and that links to students' classroom learning. We have presented in this paper a web of evidence that suggests that the DLLT is indeed valid for use in the field by those in instructional support roles to evaluate and assess effective teaching practices. In Study 1, we showed that each rubric has a high level of internal consistency, that high levels of inter-rater reliability can be achieved by in-school coaches, that the item difficulties order consistent with a hypothesized developmental trajectory, and that the DLLT can distinguish between novice and more experienced teachers at one time point. In Study 2, we showed that the DLLT is sensitive to growth in teaching practice over the course of a 3-year intervention, that this growth is related to exposure to one-on-one coaching, and that teacher practice as measured by the DLLT is related to teachers' value added to student achievement year by year.

Taken together, these results document statistical associations for DLLT measures that we would expect to see if the DLLT is functioning as a valid indicator of the quality of classroom instruction. While questions can be raised about the causal warrant associated with any one piece of analysis detailed in this paper (and no claims are made to this effect), the overall pattern of results does form a supportive evidence web about the overall reliability and validity of the DLLT.

However, there are limitations in our studies that warrant further investigation. Fundamentally, validity can only be established in a specific context of use; validity is not a general property of an instrument. For this reason, the DLLT needs to be used and tested in other contexts if it is to be used for purposes other than instructional coaching and by people other than expert literacy practitioners. Specifically, there are three contexts that could be useful applications of the DLLT and that warrant further research.

(1) Other comprehensive literacy classrooms

We developed the DLLT to align with activities typically found in a comprehensive literacy classroom; many teachers who ascribe to a comprehensive view of literacy instruction use most or all of these activities in their teaching. For this reason, we believe that the DLLT would be a useful rubric outside of the Literacy Collaborative context where it has been first tested, and have anecdotal evidence to that effect. However, more research is needed to determine if the DLLT can be used reliably in comprehensive literacy classrooms that are not part of the Literacy Collaborative program and by instructional coaches who have not had the specific Literacy Collaborative training.

(2) Use in more high-stakes applications

There is a current demand within the education field to develop tools and systems for rigorous teacher evaluation, to identify both low and high performers. In this

context, it seems like the DLLT could be a useful tool; indeed, such high-stakes evaluation of teachers could benefit from rigorous evaluations tied to both practice and student achievement. However, the DLLT would have to be tested for use in this context. Specifically, if this instrument were to be used in high-stakes evaluation, attention needs to be directed to possible sources of unreliability in data collection, since bias could easily be introduced if this tool is linked to high-stakes outcomes for teachers or their support staff.

(3) Use by individuals other than coaches

Perhaps principals, researchers, or independent evaluators might have an interest in using the DLLT, but this use would also need to be investigated. Primarily, we would need to know if individuals with less expertise in this instructional *content and pedagogy* could reliably use the DLLT. This is particularly important since some of the category descriptions are high inference. That is, they presume a body of professional knowledge shared by the observers. When such professional knowledge is shared, the results demonstrate that reasonably inter-rater reliability can be achieved. However, if such professional knowledge is not established, use of the DLLT may require further training and validation.

The benefit of the DLLT, as we have tested it, is in providing practitioners with tools for assessing teaching practice in a rigorous way that is meaningful and provides scaffolding for ongoing efforts to improve teaching. For practitioners who would like to use the tool in this way, we have several recommendations and suggestions. First, it is important that observers be well trained in the content and pedagogy of comprehensive literacy. Second, it is important to regularly check on inter-rater reliability throughout the program of use of the DLLT. This would also protect against the possibility that coaches introduce bias into their ratings, since they (implicitly or explicitly) may feel that their teachers' performance and growth might reflect on them. Finally, additional testing of the DLLT will continue to improve its applications for use; thus, we recommend continued evaluation of this tool in varied contexts. As it stands now, we feel confident that we have amassed sufficient reliability and validity evidence to support broader use in practice (and further study of the DLLT as these clinical uses occur) to assess teaching practice in K-2 comprehensive literacy classrooms.

Acknowledgements

The work described in the current article was supported by a Teacher Quality Grant from the Institute for Educational Sciences (IES), R305M040086. We are appreciative of the support provided by IES. All errors of fact, omission, and/or interpretation are solely the authors' responsibility. The research team involved in this study included affiliates of the coaching program being investigated. This collaboration informed the study design and the development of tools to measure teacher practice. The analytical team worked independently from those affiliated with the program to ensure objectivity.

Notes

1. The complete rubric can be found online (http://www.carnegiefoundation.org/files/DLLT_20120904.pdf).
2. Note that Atteberry and Bryk (2011) reported that teachers received 0.79 one-to-one coaching sessions per eligible month with a standard deviation of 0.63. In the analyses

reported on here, we use averages from an HLM model that produced an average estimate for each teacher in each year, adjusted using Empirical Bayes.

3. Field observation by study staff suggested some upward scale drift by coaches toward the end of the 1st year of coaching activity. Consequently, study staff retrained coaches during the summer of 2006, and the inflated scores reported in the spring dropped down to more realistic levels in the fall of 2006. While we include the spring 2006 data in our analyses, we add a fixed effect to control for this one-time measurement artefact.
4. In the analyses presented in this paper, we used Empirical Bayes estimates of average coaching per month in lieu of the raw coach reports. By doing so, we were able to use estimates of each individual's exposure adjusted for possible error in the coach's report.
5. Comprehensive literacy learning theory suggests that students' acquisition of reading skill is influenced by both reading and writing instruction. For this reason, the DLLT includes elements from both reading and writing instruction. The study only used standardized reading measures as outcome variables, since reliable writing tests were not available for early primary grades.

Notes on contributors

Heather J. Hough (PhD) is a Policy Fellow with the Public Policy Institute of California. Her primary focus is on improving human capital systems in K-12 education, including teacher compensation, support, and accountability. Her current research focuses on measuring the effect of district- and state-level policy interventions. Heather has a PhD in Education Policy from Stanford University and has worked in the Center for Education Policy at SRI International.

David Kerbow is a Senior Research Associate at the Urban Education Institute, University of Chicago. His research has concentrated on urban schooling, literacy assessment, student mobility, and policy evaluation. He is a co-developer of the STEP Classroom Literacy Assessment which is currently being used extensively across many urban school districts.

Anthony S. Bryk is the ninth president of the Carnegie Foundation for the Advancement of Teaching, where he is leading work on transforming educational research and development by more closely joining researchers and practitioners to improve teaching and learning. Formerly, he held the Spencer Chair in Organizational Studies in the School of Education and the Graduate School of Business at Stanford University from 2004 until assuming Carnegie s presidency in September 2008. He came to Stanford from the University of Chicago, where he was the Marshall Field IV Professor of Urban Education in the sociology department, and where he helped found the Center for Urban School Improvement, which supports reform efforts in the Chicago Public Schools. He also created the Consortium on Chicago School Research, a federation of research groups that have produced a range of studies to advance and assess urban school reform. He is a member of the National Academy of Education and was appointed by President Obama to the National Board for Education Sciences in 2010. In 2011, he was elected as a member of the American Academy of Arts and Sciences. He is one of America's most noted educational researchers. His 1993 book, *Catholic Schools and the Common Good*, is a classic in the sociology of education. His deep interest in bringing scholarship to bear on improving schooling is reflected in his later volume, *Trust in Schools*, and in the most recent book, *Organizing Schools for Improvement: Lessons from Chicago* (Chicago Press, 2009). Dr. Bryk holds a B.S. from Boston College and an Ed.D. from Harvard University.

Gay Su Pinnell is Professor Emeritus in the School of Teaching and Learning at The Ohio State University. She has extensive experience in classroom teaching and field-based research, and has developed and implemented comprehensive approaches to literacy education. She received the International Reading Association's Albert J. Harris Award for research in reading, as well as the Ohio Governor's Award for education. She also received the Charles A. Dana Foundation Award, given for pioneering contributions in the fields of health and education. She is a member of the Reading Hall of Fame. With Irene Fountas, she is co-author of *Guided Reading: Good First Teaching for All Children* (1996), *Matching Books to Readers: Using Leveled Books in Guided Reading, K-3* (1999), *Word Matters: Teaching Phonics*

and Spelling in the Reading/Writing Classroom (1998), *Help America Read: A Handbook for Volunteers* (1997), *Interactive Writing: How Language & Literacy come Together, K-2* (2000), and *Guiding Readers & Writers, Grades 3-6* (2000). She also has co-authored *Systems for Change: A Guide to Professional Development*, with Carol Lyons. With Irene Fountas, she also developed a four-volume set, *Phonics Lessons: Letters, Words, and How They Work*, for kindergarten, first, second, and third grades, as well as an early literacy intervention program (Leveled Literacy Intervention) that includes student books. Her recent publications, co-authored with Fountas, are *Teaching for Comprehending and Fluency* (2007), *When Readers Struggle: Teaching that Works* (2009), *The Continuum of Literacy Learning: Pre-K-8* (2910), and *Genre Study: Teaching with Fiction and Nonfiction Books* (2012).

Emily Rodgers is an associate professor in the College of Education and Human Ecology at The Ohio State University. She has worked in schools as a reading specialist and special education teacher. Her research focuses on the professional development of teachers and scaffolding literacy learning particularly for young children having great difficulty learning to read and write.

Emily Dexter, Ed.D., is an independent researcher and educational consultant in the greater Boston area. During the time that this research was being conducted, she was the Director of Research for the Center for Reading Recovery and Literacy Collaborative at Lesley University. Her training is from the Harvard Graduate School of Education, where she received her doctorate in Human Development in 2000. She is the co-author of Literacy and Mothering: *How Women's Schooling is Changing the Lives of the World's Children* (Oxford University Press, 2012), which was awarded the 2013 Eleanor Maccoby Award from the American Psychological Association for its contribution to the field of developmental psychology.

Carrie Hung is a doctoral candidate and graduate research associate at the Ohio State University.

Patricia L. Scharer is a Professor of Education and Human Ecology at The Ohio State University. She is actively involved in two literacy projects Literacy Collaborative, a K-6 school reform model training on-site coaches to support teacher professional development, and Reading Recovery, a research-based intervention for first-grade students experiencing difficulty learning to read and write. In 2004, Dr. Scharer collaborated with colleagues at both OSU and the University of Chicago to receive a $3,914,000 Federal Teacher Quality grant to study the impact of Literacy Collaborative professional development on both teachers and student's literacy achievement. In October, 2010, Dr. Scharer and colleagues Dr. Jerry D'Agostino and Dr. Emily Rodgers were awarded a $54 million federal i3 grant to scale-up Reading Recovery across the US. Dr. Scharer's research interests include early literacy development, phonics and word study, and the role of children's literature to foster both literary development and literacy achievement. Her research has been published in *Reading Research Quarterly*, *Research in the Teaching of English*, *Educational Leadership*, *Language Arts*, *The Reading Teacher*, *Reading Research and Instruction*, *Journal of Reading Recovery*, *Literacy Teaching & Learning*, and the yearbooks of the National Reading Conference and the College Reading Association.

Irene C. Fountas, a Professor in the Graduate School of Education at Lesley University in Cambridge, Massachusetts, has been a classroom teacher, language arts specialist, and consultant in school districts across the nation and abroad. She works extensively in the literacy education field and directs the Center for Reading Recovery and Literacy Collaborative at Lesley University.

References

American Federation of Teachers. (2010). Continuous improvement: Making evaluation a tool for increasing teacher and student learning. *American Educator*, *34*(2), 36–41.

Armbruster, B.B., Lehr, F., & Osborn, J. (2001). *Put reading first: Kindergarten through grade 3*. Retrieved from http://www.nifl.gov/publications/pdf/PRFbooklet.pdf

Atteberry, A., & Bryk, A.S. (2011). Analyzing teacher participation in literacy coaching activities. *The Elementary School Journal*, *112*, 356–382.

Ball, C., & Gettinger, M. (2009). Monitoring children's growth in early literacy skills: Effects of feedback on performance and classroom environments. *Education and Treatment of Children*, *32*, 189–212.

Ball, D.L., & Rowan, B. (2004). Introduction: Measuring instruction. *The Elementary School Journal*, *105*, 3–10.

Beck, I., & McKeown, M. (2001). Text talk: Capturing the benefits of read-aloud experiences for young children. *The Reading Teacher*, *55*, 10–20.

Biancarosa, G., Bryk, A.S., & Dexter, E.R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal*, *111*, 7–34.

Bill and Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project* (MET Project Research Paper). Seattle, WA: Author.

Bitter, C., O'Day, J., Gubbins, P., & Socias, M. (2009). What works to improve student literacy achievement? An examination of instructional practices in a balanced literacy approach. *Journal of Education for Students Placed at Risk*, *14*, 17–44.

Bransford, J.D., Brown, A., & Cocking, R. (1999). *How people learn: Mind, brain, experience, and school*. Washington, DC: National Academies Press.

Cazden, C. (2001). *Classroom discourse* (2nd ed.). Portsmouth, NH: Heinemann.

Clay, M. (1979). *Reading: The patterning of complex behavior*. Auckland, New Zealand: Heinemann Educational Books.

Clay, M. (2001). *Change over time in children's literacy development*. Portsmouth, NH: Heinemann.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision & Curriculum Development.

Doherty, R.W., Hilberg, R.S., Epaloose, G., & Tharp, R. (2002). Standards performance continuum: Development and validation of a measure of effective pedagogy. *The Journal of Educational Research*, *96*, 78–89.

Edmunds, M., & Briggs, K. (2003). The instructional content emphasis instrument: Observations of reading instruction. In S. Vaughn & K. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 31–52). Baltimore, MD: Paul H. Brooks.

Estrada, P. (2004). Patterns of language arts instructional activity and excellence in fourth-grade culturally and linguistically diverse classrooms. In H.C. Waxman, R.G. Tharp, & R.S. Hilberg (Eds.), *Observational research in US classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 122–143). Cambridge, UK: Cambridge University Press.

Flyvbjerg, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge, UK: Cambridge University Press.

Foorman, B., & Schnatschneider, C. (2003). Measurement of teaching practices during reading/language arts instruction and its relationship to student achievement. In S. Vaughn & K. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 1–30). Baltimore, MD: Paul H. Brooks.

Fountas, I.C., & Pinnell, G.S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann.

Fountas, I.C., & Pinnell, G.S. (2006). *Teaching for comprehending and fluency: Thinking, talking, and writing about reading, K-8*. Portsmouth, NH: Heinemann.

Good, R.H., & Kaminski, R.A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Goswami, U. (1999). Causal connections in beginning reading: The importance of rhyme. *Journal of Research in Reading*, *22*, 217–254.

Grace, C., Bordelon, D., Cooper, P., Kazelskis, R., Reeves, C., & Thames, D.G. (2008). Impact of professional development on the literacy environments of preschool classrooms. *Journal of Research in Childhood Education*, *23*, 52–81.

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.

Greenwood, C.R., Abbott, M., & Tapia, Y. (2003). Ecobehavioral strategies: Observing, measuring, and analyzing behavior and reading interventions. In S. Vaughn & K. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 53–82). Baltimore, MD: Paul H. Brooks.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores*. Cambridge, MA: National Bureau of Economic Research.

Horn, M., & Giacobbe, M. (2007). *Talking, drawing, writing: Lessons for our youngest writers*. Portland, ME: Stenhouse.

Juel, C., & Minden-Cupp, C. (2000). Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly*, *35*, 458–492.

Kane, T.J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2010). *Using student performance data to identify effective classroom practices*. Cambridge, MA: National Bureau of Economic Research.

Linacre, M. (2010). *Misfit diagnosis: Infit outfit mean-square standardized?* Retrieved from http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm

Matsumura, L.C., Garnier, H.E., & Slater, S.C. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, *13*, 267–300.

Matsumura, L.C., Slater, S.C., Junker, B., Peterson, M., Boston, M., Steele, M., & Resnick, L. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the instructional quality assessment* (CSE Technical Report 681). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

McCarrier, A., Pinnell, G.S., & Fountas, I.C. (2000). *Interactive writing: How language & literacy come together, K-2*. Portsmouth, NH: Heinemann.

McGraw-Hill. (2001). *Terra Nova technical quality: Reliable, useful results based on psychometric excellence*. Monterey, CA: CTB McGraw-Hill.

Morrow, L. (2008). *Literacy development in the early years: Helping children read and write* (6th ed.). Needham Heights, MA: Allyn & Bacon.

National Governors Association Center and Council of Chief State School Officers. (2010). *The common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/about-the-standards/key-points-in-english-language-arts

Neufeld, B., & Roper, D. (2003). *Coaching: A strategy for developing instructional capacity*. Washington, DC: Aspen Institute.

Pentimonti, J.M., & Justice, L.M. (2010). Teachers' use of scaffolding strategies during read alouds in the preschool classroom. *Early Childhood Education Journal*, *37*, 241–248.

Pinnell, G.S., & Fountas, I.C. (1998). *Word matters: Teaching phonics and spelling in the reading/writing classroom*. Portsmouth, NH: Heinemann.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417–458.

Rumelhart, D. (1994). Toward an interactive model of reading. In R.B. Ruddell, M.R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed., pp. 864–894). Newark, DE: International Reading Association.

Sanders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Smith, M., Dickinson, D., Sangeorge, A., & Anastasopoulos, L. (2002). *Early language and literacy classroom observation toolkit, research edition*. Baltimore, MD: Brookes.

Taylor, B.M., Pearson, P.D., Peterson, D.S., & Rodriguez, M.C. (2005). The CIERA school change framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly*, *40*, 40–69.

Tharp, R., & Gallimore, R. (1991). *The instructional conversation: Teaching and learning in social activity*. Retrieved from Center for Research on Education, Diversity & Excellence, UC Berkeley: http://escholarship.org/uc/item/5th0939d

Tyler, J.H., Taylor, E.S., Kane, T.J., & Wooten, A.L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review*, *100*(2), 256–260.

Wolf, M.K., Crosson, A.C., & Resnick, L.B. (2006). *Accountable talk in reading comprehension instruction* (CSE Technical Report 670). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis. Rasch measurement*. Chicago, IL: MESA Press.
Wright, B.D., & Stone, M.H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Appendix 1. Rubric example: Writing Workshop

Writing Workshop	**Date: _____**	**Time began: _____**	**Time ended: _____**	**☐ Teacher never uses this element**
Minilesson: The teacher:				
__May give directions on what to do during writing but there is no explicitly stated principle (for example, gives a "story starter" or worksheet assignment).	__Provides a minilesson but the principle is not clearly stated or receives brief attention.	__Provides a minilesson with a principle stated in partially clear language.	__Provides a minilesson that is clearly stated and focused on a writing principle.	
__Provides no demonstration or understandable example of the principle; primarily shows students how to complete an assignment.	__Provides limited demonstration or example of the use of the principle or its application to writing.	__Provides some demonstration or example in brief form that shows how the principle is used in writing.	__Provides a clear and explicit demonstration or example of what students need to learn as writers (craft, conventions, or process).	
__ Does not check on students' understanding of how to apply the lesson to their own writing.	__ Briefly checks on students' understanding of principle or application but with limited interaction from students.	__ Checks on students' understanding of principle or application with some interaction from students but there is evident potential for misunderstanding and no evidence in students' comments that they understand.	__ Checks on understanding of principle or application and elicits comments from students that are evidence of understanding.	

(*continued*)

Appendix 1. (*Continued*)

Writing Workshop	**Date: _____**	**Time began: _____**	**Time ended: _____**	**☐ Teacher never uses this element**
Writing and Conferring: The teacher: (Some teachers may be working with small groups of children. In this case, consider these "group conferences" and do the rating across both individual and group conferences.)				
__ May check with some students to monitor progress, but does not confer with students about their writing.	__Confers with some students and discusses their writing; most interactions are general and not explicitly focused on helping students learn about the writing process.	__Confers with several students; interactions may improve individual pieces but few conferences are focused on helping students learn about the **writing process**.	__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 **writing process**.	
__Does not take notes.	__Takes occasional notes; not evident that sufficient information is being recorded.	__Takes notes in most conferences about individual students' writing.	__There is consistent evidence of note taking and continuity from previous conferences.	
Sharing: The teacher:				
__Provides no opportunity for sharing. Students do not comment about each other's writing.	__Provides time for students to share their writing but there are limited or very general comments from students about each other's writing.	__Provides time for students to share their writing; students comment about each other's writing with some specific detail.	__ Provides time for students to share their writing; students comment specifically about other students' writing and show understanding of strategies or craft of writing.	
__ Does not provide opportunity for sharing or make helpful comments about students' writing.	__ Makes some comments to encourage or praise student writing but does not reinforce the principle or strategies for writing during sharing.	__ Makes some specific comments about writing shared by students but does not tie comments to the principle or strategies for writing.	__ Makes explicit and helpful comments about writing shared by students and clearly reinforces the principle and strategies for writing.	

Appendix 2. Measurement statistics.

Table 2–1. Read Aloud.

		Item Difficulty	*Infit Statistic*
RA3	Discussion after reading efficiently builds on overall meaning and extends students' thinking about the text.	0.71	0.95
RA2	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
RA1	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

Note: The descriptors for the rubric items used in this appendix are the Category 4 (expert practice) descriptors for each respective item. The estimation of the item difficulty and infit statistics are actually based on all of the observational data reported from all four categories on each rubric item. For more details about the estimation of item difficulty and infit statistics, see Wright and Masters (1982).

Table 2–2. Shared Reading.

		Item Difficulty	*Infit Statistic*
SR3	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
SR2	Teacher engages almost all children in active shared reading of the text.	−0.33	1.05
SR1	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

Table 2–3. Guided Reading.

		Item Difficulty	*Infit Statistic*
GR6	Makes superbly chosen, specific teaching points that help students engage in effective processing of text.	1.23	1.12
GR5	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
GR7	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
GR4	Samples oral reading and demonstrates, reinforces, and consistently prompts (as needed) for effective reading behaviors and problem solving actions.	−0.08	0.75
GR2	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
GR3	Engages students in a conversation that brings them into the text and supports thinking about the meaning of the text.	−0.64	0.99
GR1	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

Table 2–4. Interactive Writing.

		Item Difficulty	*Infit Statistic*
IW5	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
IW3	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
IW1	Engages children in interesting experiences and a rich and purposeful discussion before writing;	0.21	0.88
IW4	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
IW2	Makes writing a highly purposeful and connected activity.	−1.38	1.27

Table 2–5. Writing Workshop.

		Item Difficulty	*Infit Statistic*
WW5	There is consistent evidence of note taking and continuity from previous conferences.	2.12	2.07
WW6	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
WW7	Makes explicit and helpful comments about writing shared by students and clearly reinforces the principle and strategies for writing.	−0.05	1.03
WW1	Provides a mini-lesson that is clearly stated and focused on a writing principle.	−0.50	0.77
WW2	Provides a clear and explicit demonstration or example of what students need to learn as writers (craft, conventions, or process).	−0.58	0.73
WW3	Checks on understanding of principle or application and elicits comments from students that are evidence of understanding.	−0.58	0.83
WW4	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 *writing process*.	−0.89	0.97

Table 2–6. Word Study.

		Item Difficulty	*Infit Statistic*
WS7	Students actively participate in sharing, comment on their work, and show evidence of learning the principle.	1.83	1.20
WS6	Teacher clearly restates the principle and reinforces learning, through examples of students' work.	1.52	1.15
WS4	Provides an application task that is appropriate and has strong potential for helping students develop greater understanding of the principle.	0.79	0.93
WS3	Clearly demonstrates and explains the application task and explicitly relates it to the principle.	−0.06	0.65
WS1	Provides a mini-lesson with a clearly and explicitly stated principle *or* asks children to derive the principle from examples and to state the principle clearly and explicitly.	−0.38	0.86
WS2	Uses good examples; teacher checks for understanding and helps students understand how the principle is related to reading and writing.	−0.98	0.70
WS5	Explains the application task in a way that enables almost all students to perform the task independently.	−1.15	1.12

Table 2–7. General Aspects of Teaching.

		Item Difficulty	*Infit Statistic*
GAT8	*Quality of Interactions:* Student discussion builds on the comments of other students; students provide evidence to support their ideas based on the text.	1.50	0.94
GAT4	*Student Engagement:* 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
GAT5	*Student Engagement:* Transitions are orderly and efficient.	0.39	1.00
GAT7	*Quality of Interactions:* Students have many opportunities to talk to, and learn from, each other.	0.03	1.17
GAT9	*Sense of Community:* 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
GAT3	*Classroom Materials and Organization:* Student/teacher generated charts are accessible, relevant and routinely used by teacher and students to guide learning.	−0.02	1.20
GAT2	*Classroom Materials and Organization:* Organization works for maximum student independence; use and placement of materials in the classroom is obvious.	−0.59	0.85
GAT1	*Classroom Materials and Organization:* Materials are highly organized for efficient use by the teacher and students.	−0.71	1.00
GAT6	*Quality of Interactions:* The teacher consistently listens and responds to students.	−1.03	0.98

Table 2–8. Teaching for Strategies.

		Item Difficulty	*Infit Statistic*
TS9	*Teaching for Fluency and Phrasing:* Across reading instruction, teacher demonstrates, attends to, reinforces, and prompts for reading that is fluent, phrased, and well stressed.	0.96	1.30
TS5	*Teaching for Inferential and Analytic Thinking:* 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	0.93
TS6	*Teaching for Word Solving:* 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
TS3	*Teaching for Inferential and Analytic Thinking:* 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
TS7	*Teaching for Word Solving:* 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
TS10	*Teaching for Fluency and Phrasing:* Teacher assists children when there is evidence of dysfluent reading in various contexts; teacher avoids interrupting fluent reading.	0.03	1.51
TS8	*Teaching for Word Solving:* 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
TS2	*Teaching for Literal Thinking:* Teacher helps students learn how to search for and use information that is in the text.	−0.38	0.91
TS4	*Teaching for Inferential and Analytic Thinking:* 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
TS1	*Teaching for Literal Thinking:* 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

Appendix 3. Full model results from Study 2

	Baseline Model				*Growth Model*				*Growth Model with Coaching*			
FIXED EFFECT	*Coef.*	*se*	*t Ratio*	*p Value*	*Coef.*	*se*	*t Ratio*	*p Value*	*Coef.*	*se*	*t Ratio*	*p Value*
Intercept	0.411	0.157	2.61	0.014	−0.293	0.169	−1.73	0.083	−0.278	0.169	−1.64	0.100
Growth					0.161	0.031	5.25	<0.001	0.147	0.03	4.58	<0.001
Monthly coaching									0.134	0.05	2.74	0.007
RANDOM EFFECT	*Var.*	*df*	X^2	*p Value*	*Var.*	*df*	X^2	*p Value*	*Var.*	*df*	X^2	*p Value*
Level 1 & 2	0.854	1079	12834.93	<0.001	0.315	1078	4763.27	<0.001	0.316	1078	4770.43	<0.001
Teacher level												
Intercept	0.695	199	999.08	<0.001	0.778	195	617.47	<0.001	0.767	195	615.63	<0.001
Growth					0.001	195	264.19	0.001	0.001	194	262.72	0.001
School level												
Intercept	0.296	15	88.21	<0.001	0.341	15	80.76	<0.001	0.342	15	82.36	<0.001
Growth					0.014	15	211.51	<0.001	0.015	15	211.16	<0.001