Learning About Teachers' Literacy Instruction From Classroom Observations

Ben Kelcey
University of Cincinnati, Ohio, USA

Joanne F. Carlisle
University of Michigan, Ann Arbor, USA

ABSTRACT

The purpose of this study is to contribute to efforts to improve methods for gathering and analyzing data from classroom observations in early literacy. The methodological approach addresses current problems of reliability and validity of classroom observations by taking into account differences in teachers' uses of instructional actions (e.g., modeling) in specific skill areas (e.g., fluency, reading comprehension). The findings from observations of second- and third-grade teachers' literacy instruction showed that teachers' instructional actions differed by literacy skill area and were more consistent within than across skill areas. Furthermore, teachers' uses of instructional actions in a given skill area were more strongly associated with students' gains in achievement in that skill area than were teachers' uses of actions across all skill areas. The approach offers significant improvements in methods to identify features of effective literacy instruction.

In recent years, the drive to improve students' achievement in literacy has focused interest on how teachers teach and what characteristics of their teaching are associated with students' literacy achievement. Although there are a variety of ways to gather information about teachers' literacy instruction (e.g., student surveys, gains on state tests), many consider observation of instruction the most promising approach because it offers the opportunity for teachers and researchers to gain insights into effective literacy instruction (e.g., Hoffman, Maloch, & Sailors, 2011; MET Project, 2012, 2013). Because of the promise of this method for studying learning in classrooms, observation systems have been used to examine a wide range of factors that might affect teachers' instruction as it relates to students' academic achievement—for example, the quality of the climate in classrooms (e.g., Pianta & Hamre, 2009), teachers' sensitivity to student needs (e.g., Connor et al., 2009), and the texts that are used in literacy lessons (Hoffman, Sailors, Duffy, & Beretvas, 2004).

Despite the potential of classroom observations to identify differences in instructional actions of more and less effective teachers, there are reasons to be concerned about the extent to which observation studies can yield measures of effective teaching. Researchers have expressed concerns about the lack of rigor in the design and analytic methods used in many such studies (Hoffman et al., 2011). Two particularly pressing problems are the low levels of reliability in descriptions of teachers' instruction and the low external validity of such descriptions in predicting students' literacy achievement. These problems were noted in the report of the Measures of Effective Teaching project; even with as many as four observations and observers, most classroom observation scores demonstrated low reliability (e.g., on the order of .5) and

Reading Research Quarterly, 48(3)
pp. 301–317 | doi:10.1002/rrq.51

were minimally associated with student achievement (MET Project, 2012). When studies fall shy of high standards of reliability and validity, we need to question whether they accurately portray aspects of literacy instruction that contribute to student achievement.

One factor that contributes to low reliability is the variation in the ways teachers teach different lessons. For example, the differences in instruction among lessons within teachers has been found to account for as much as a quarter of the total variation in observed teaching (Carlisle, Kelcey, Berebitsky, & Phelps, 2011; Hill, Charalambous, & Kraft, 2012; MET Project, 2012). Although such variation is to be expected given the wide range of knowledge and skills teachers seek to impart, the sheer magnitude of lesson-to-lesson variation has significant potential to undermine researchers' power to identify features of effective literacy instruction. In analyses of classroom instruction data, lesson-to-lesson variation makes it difficult to differentiate among teachers and identify effective features because this variation weakens the ability to reliably infer whether observed differences are indicative of systematic differences that persist across most lessons or whether observed differences are atypical and limited to a specific lesson.

Although findings of lesson-to-lesson variability are not surprising (Stodolsky, 1990), what is surprising is that extant research has tended to treat lesson-to-lesson differences in instruction as if they represented random noise or error to be averaged over. Our proposal is that differences among lessons within a teacher are, in part, a function of the literacy skill area (e.g., fluency, vocabulary). For this reason, data from observation studies of literacy instruction should not be collapsed across lessons in different literacy skill areas; rather, methods of data analysis should retain differences in the ways teachers teach skills in different skill areas.

To determine whether this approach leads to a more reliable and valid picture of literacy instruction, our observational study examines how second- and third-grade teachers teach in different literacy skill areas (i.e., the instructional actions they choose to use in lessons in different skill areas). Our analysis of classroom observational data in this study involves tracking both within- and between-teacher variation, while allowing for systematic differences among lessons taught in different skill areas. The first goal is to determine the extent to which there is evidence that teachers' uses of instructional actions differ by literacy skill area and are less variable within than across skill areas. The second goal is to investigate the validity of the approach by examining the extent to which teachers' uses of instructional actions within a given skill area are more related to students' progress in literacy achievement in that area than is their use of instructional actions across all literacy skill areas.

In what follows, we discuss issues of variability in teaching and how these have been addressed in previous studies. We explain why we expect teachers' instruction to be less variable within literacy skill areas (e.g., within different lessons on phonics) than across literacy skill areas (e.g., collapsed across lessons in phonics, fluency, comprehension, and other skill areas). Then, we provide an overview of the three dimensions of instruction (and instructional actions that represent these dimensions) that we used to examine teachers' early literacy instruction. Our approach adheres to the guidelines for rigorous study of classroom instruction, described by Hoffman and colleagues (2011) as a "careful mapping between theory, data collection and data analysis" (p. 17).

Variability in Instruction

How teachers teach in a given lesson reflects the nature of the concepts, skills, and knowledge that students need to learn in that lesson. Thus, teachers are bound to teach lessons in different areas and topics in different ways (e.g., Barr & Dreeben, 1983; Stodolsky, 1990). This explains why previous observational studies have found little stability in the instructional actions teachers use to deliver content to the students; there is similar variability in other aspects of their instruction, such as their choice of materials (Carlisle et al., 2011; Correnti & Rowan, 2007).

There are reasons to expect that we should find more stability in teachers' uses of instructional actions if we examine how they teach literacy in specific skill areas (e.g., vocabulary, fluency) instead of across lessons of different skill areas. One reason is that a comparison of studies of effective instruction in specific literacy skill areas reveals noteworthy differences; that is, instructional actions tend to reflect the nature of the skills and knowledge that students need to acquire in a given area. For example, the National Reading Panel (National Institute of Child Health and Human Development, 2000) carried out analyses of studies in five literacy skill areas (i.e., phonemic awareness, phonics, vocabulary, fluency, reading comprehension). Examining their findings, we note that effective instruction in fluency emphasizes the important role of practice (e.g., opportunities to read texts aloud or with repeated readings), whereas effective vocabulary instruction emphasizes strategy instruction (e.g., how to estimate a word's meaning from context). Another reason is that students' progress in reading is often assessed by skill area; for example, in the early elementary years, teachers are likely to monitor students' progress in word reading, vocabulary, oral reading (fluency), composition writing, and reading comprehension.

Despite recognition that there are important sources of variability across lessons within teachers, researchers studying literacy instruction have commonly described teachers' uses of particular instructional actions by summing or averaging instances of instructional actions across all lessons and skill areas (e.g., Foorman et al., 2006). For example, Taylor, Pearson, Peterson, and Rodriguez (2003, 2005) observed teachers' literacy instruction as part of a study of elementary school reform. They used the observation system CIERA Classroom Observation Scheme to collect data on an array of features of literacy lessons (19 in all), such as teachers' uses of instructional actions (e.g., telling) and their classroom organization (e.g., small group). Their results provided descriptive information about how frequently teachers used different instructional actions, and were collapsed across teachers, days, and skill areas. Results indicated that, for example, telling (i.e., teachers' explanations) occurred in 60% of the observation segments.

In other studies, averages are used to represent the time teachers allotted to particular aspects of instruction. For example, Piasta, Connor, Fishman, and Morrison (2009) reported that the average first-grade student received 7.43 minutes of explicit decoding. As another example, Foorman and her colleagues (2006) reported that feedback was observed on average less than 2% of the observation time in the two districts participating in their study. This averages does not take into account the skill areas in which feedback was observed. For example, we do not know whether teachers provided feedback more often in phonics than in other literacy skill areas (e.g., writing).

The practice of collapsing data across literacy skill areas runs counter to the generally accepted premise that teachers are likely to teach lessons in different skill areas in somewhat different ways. Collapsing data across skill areas can also inadvertently introduce unsupported assumptions. From a methodological standpoint, an important assumption underlying the soundness of approaches that collapse features of observed instruction across skill areas is the concept of exchangeability (e.g., Webb, Shavelson, & Haertel, 2006). Applied to classroom observations of literacy instruction, exchangeability assumes that instruction in different skill areas is interchangeable and identical across skill areas (e.g., a teacher uses a set of instructional actions with the same frequency across skill areas).

There are two significant problems with collapsing measures of observed instruction across skill areas when skill areas are not interchangeable (i.e., when teachers' uses of instructional actions differ across skill areas). The first concerns the reliability with which we can describe teachers' instruction. For the purposes of differentiating more and less effective instruction, within-teacher variation introduces measurement error (i.e., unreliability) because it weakens our ability to identify which parts of a teacher's observed instruction are likely to have occurred in all of his or her lessons. Because there is likely to be greater variability across lessons in different skill areas than across lessons within the same skill area, the reliability with which we can describe teachers' instruction will generally be lower. In turn, descriptions that characterize literacy instruction by collapsing across skills areas have less ability to differentiate among features of instruction and less power to identify which features of instruction are effective.

The second problem concerns the validity of indexes formed through classroom observations. Even given sufficiently reliable measures, the underlying meaning and predictive validity of indexes formed by averaging across different skill areas can be distorted if instruction differs by skill area. Of particular relevance here is the risk of construct underrepresentation (Messick, 1989), which can occur when measures fail to adequately discriminate among levels of a construct (e.g., dimension of instruction). A common source of construct underrepresentation noted in validity studies in other areas arises when methods rely too heavily on general aspects of a construct (e.g., overall averages) that do not fully capture the construct in ways that discriminate among instructional profiles of teachers (Klein & Stecher, 1998).

For a hypothetical example, let us assume that regular use of explicit instruction in phonics lessons is associated with higher levels of student achievement in phonics but that explicit instruction is negatively associated with student achievement in reading comprehension lessons. In correlating student achievement in phonics to teachers' average use of explicit instruction across lessons in both skill areas, we may find no association because teachers who regularly engage in explicit instruction in phonics lessons but not comprehension lessons will be presented as the same as those who regularly use explicit instruction in comprehension but not phonics lessons.

Because both theory and empirical findings (as reviewed previously) suggest that teachers' instruction varies by literacy skill area, we designed a study to challenge an approach to data analysis that assumes that instruction in different skill areas is exchangeable. Specifically, we expected to find that teachers' instruction systematically varied by skill area and thus was less variable and more reliable within than across skill areas. Similarly, there are reasons to suspect that indexes formed by collapsing across lessons in different skill areas may underrepresent teachers' instruction because the indexes fail to sufficiently capture the variability inherent in teaching different skill areas. A close association between the area of literacy instruction and the

content focus of a student achievement test should yield more interpretable findings concerning effective practice than would situations in which what is taught and what is tested have little direct connection (Shavelson, Webb, & Burstein, 1986). Thus, we also expected that instruction within skill areas would be more predictive of student achievement than it would if we were to collapse data across skill areas.

Dimensions of Instruction

In commenting on descriptions of different classroom observation systems, Douglas (2009) noted the importance of using a theoretical model that looks at components or dimensions of classroom instruction known or thought to contribute to effective instruction. Because our goal is to examine how teachers teach early literacy, we focus on dimensions that have been found to be central to the process of teaching. Influenced by Shulman (1987) and Roehler and Duffy (1991), we are interested in how teachers talk, show, enact, or otherwise represent information or ideas so their students acquire a deeper knowledge of literacy and the skills they need to become proficient readers and writers.

In developing the observational system, we selected three theoretical dimensions identified by previous researchers as contributing to effective instruction (e.g., Brophy & Good, 1986; Good & Mulryan, 1990; Hoffman, 1991; Roehler & Duffy, 1991; Rosenshine, 1995; Rosenshine & Stevens, 1984; Seidel & Shavelson, 2007). The dimensions and the observable instructional actions that represented these dimensions in our automated classroom observation system for reading (ACOS-R) are shown in Table 1. Although the dimensions we chose to investigate are among important aspects of teaching, we acknowledge that other researchers might choose to include other dimensions; the purpose of a given study should guide researchers in the choice of dimensions they choose to study.

The first dimension, organizing, reflects the view that effective pedagogy requires that teachers provide structure and organization for their students so they know what they are doing and why. Organizing as used in our study is akin to Cameron and Morrison's (2011) construct, teacher orienting, defined as "explanations and demonstrations about the procedures and rationale behind activities" (p. 620). Organizing refers to actions that teachers take to provide pedagogical structure to literacy lessons—that is, to help students understand the purpose and benefits of a given lesson. Porter and Brophy (1988) argued that effective teachers are clear about what they hope to accomplish through their instruction and communicate the purpose of lessons to their students. Cameron, Connor, and Morrison (2005) reported that teachers' organization of the class for assignments and the clarity of their lesson objectives was related to first graders' achievement. Our observational protocol included four instructional actions that represented organizing, such as explaining the purpose and explaining the value of the lesson (as shown in Table 1).

The second dimension, delivering literacy content, reflects the need for the teacher to choose ways to convey the content that will ensure learning in each lesson. This construct draws on previous research on teachers' choice of methods to bring the content to the students, such as explaining, coaching, or demonstrating (e.g., Foorman & Torgesen, 2001; Roehler & Duffy, 1991; Taylor et al., 2003). Similar constructs are explicit instruction (Foorman & Torgesen, 2001) and teacher-managed instruction (e.g., Connor, Morrison, & Petrella, 2004). Research suggests that beginning readers need to have the teacher explain new ideas and information, to show them what good readers do, and to give them opportunities to practice with guidance (e.g., Foorman & Connor,

TABLE 1
Instructional Actions in the Three Dimensions in the Automated Classroom Observation System for Reading

Theoretical dimension	Instructional action
Organizing (i.e., providing pedagogical structure)	• Explaining the purpose of the lesson • Explaining the value/relevance of the lesson • Giving directions for an activity • Providing a wrap-up or summary of what has been accomplished
Delivering literacy content (i.e., directing knowledge and skill acquisition)	• Telling • Modeling • Asking questions to check or mediate student learning • Providing practice or review activities
Supporting student learning (i.e., fostering their engagement and self-understanding)	• Fostering discussion • Assessing students' work; providing feedback • Giving students an opportunity to ask questions

2011). Studies have shown that the nature and amount of guided instruction provided by the teacher is related to students' language and literacy development (e.g., Duffy, Roehler, & Rackliffe, 1986; Snow, Burns, & Griffin, 1998). Our observation protocol included four instructional actions that teachers use to deliver literacy content, including telling or providing review and practice (as shown in Table 1).

The third dimension, supporting student learning, is based on theory and evidence that students learn best when they are actively involved in their learning and given feedback about their progress and performance. Effective teachers use instructional actions to promote students' interest and help them contribute to their own literacy development; these teachers are skilled at motivating their students and attending to their needs (e.g., Guthrie, 2004; Pressley, Wharton-McDonald, Raphael, Bogner, & Roehrig, 2002). Porter and Brophy (1988) stated, "Effective teachers continuously monitor their students' understanding of presentations and responses to assignments. They routinely provide timely and detailed feedback, but not necessarily in the same ways for all students" (p. 82). The observation protocol included three different ways that teachers might engage students' interest in the content and in improving their own literacy (see Table 1).

These dimensions provide a basis for our analysis of teachers' literacy instruction within and across skill areas of literacy. To develop hypotheses about possible differences in how teachers teach in specific skill areas of literacy, we turned to Rosenshine (1995), who suggested that variation in methods of instruction in different domains or skill areas might reflect the extent to which they are well structured. Well-structured lessons tend to have predictable elements, whereas less well-structured lessons are more variable in terms of what and how teachers teach. For example, in phonics, lessons might be well structured, involving a few commonly used instructional actions, such as modeling, asking questions for evaluation, and providing practice (e.g., Brady, 2011; National Institute of Child Health and Human Development, 2000), whereas lessons in reading comprehension might be less well structured, thus involving a wide variety of instructional actions (e.g., telling, modeling, providing practice, fostering discussion; e.g., Shanahan et al., 2010; Snow, 2002).

The complexities of the learning goals in specific areas might affect teachers' choice of instructional actions. For example, if fluency lessons are regularly treated as time to practice reading, both the teacher and students may rely on established procedures and everyday activities; as a result, teachers might not feel a need to explain the purpose or value of each fluency lesson (e.g., Rasinski, Homan, & Biggs, 2009). In areas such as reading comprehension, in which the content and skills are complex and varied (Shanahan et al., 2010), teachers presumably need to explain the purpose of a lesson so students understand what they are learning and why (suggesting engagement in the organizing dimension). In contrast, teachers might use different actions in teaching comprehension (e.g., actions in the delivering literacy content dimension) because of the need to model or demonstrate a procedure or manner of reasoning (Shanahan et al., 2010). Furthermore, teachers might engage more in actions to support student learning during comprehension lessons, such as fostering discussion, than in lessons focused on other literacy skill areas. Thus, there might be more variety in teachers' instructional actions in reading comprehension lessons than in other literacy skill areas.

Research Questions

The purpose of this study is to investigate the extent to which we might gain a more reliable and valid description of early literacy instruction by analyzing teachers' instruction within instead of across literacy skill areas. The investigation focuses on four research questions, each serving as an important test of the value of studying instruction by skill area. The first question tests the extent to which teachers' instruction is less variable within than across literacy skill areas (i.e., collapsing across all literacy lessons). If teachers' instruction is less variable within than across literacy skill areas, area-specific indexes are likely to be more reliable and better able to capture variation in teachers' instruction as it relates to students' achievement.

The second question is somewhat contingent on a positive answer to the first question: If teachers' instruction is less variable within than across skill areas, in what ways does instruction tend to differ by skill area? The third question examines the extent to which teachers' instruction in a given skill area contributes to gains in that area. The final research question contrasts the extent to which skill area–specific measures of instruction demonstrate stronger associations with students' achievement than do measures of instruction that are collapsed across skill areas.

Method

Sample

Our investigation focused on reading lessons in 87 second- and third-grade classrooms selected from 19 Reading First schools located in six school districts in Michigan. To qualify for participation in the Reading First program, participating school districts had to meet criteria for high levels of poverty and chronic

underachievement in reading (U.S. Department of Education, 2002). Of the 87 teachers, 44 taught second grade, and 43 taught third grade; 19% were non-White, and 11% had a master's degree in reading. On average, the teachers had 13 years of teaching experience. Classrooms averaged 23 students, of which roughly 45% were minority, 21% were in special education, and over three quarters were eligible for free or reduced-price lunch.

Measures

Iowa Tests of Basic Skills (ITBS)

As a measure of current and prior student achievement, we drew on three of the ITBS reading subtests. The reading comprehension subtest requires students to select responses to questions that follow short passages. The word analysis subtest asks students to identify and match sounds and spelling elements of words. The final subtest, vocabulary, tests students' breadth of vocabulary (e.g., relating words to their definitions). Test reliability for each subtest in both grades 2 and 3 exceeds .85 (computed with the Kuder–Richardson Formula 20; Hoover et al., 2003).

ACOS-R

The ACOS-R was designed to study teachers' instruction in elementary literacy. Observations were carried out four times a year for the duration of the teachers' literacy block, which typically lasted from 90 to 120 minutes (Carlisle et al., 2011). ACOS-R observations were carried out using a computerized tablet, which was programmed to present the protocol for coding. The categories in the coding system include not only instructional actions but also other variables that might contribute to the nature of teachers' instruction, such as the literacy skill area of the lesson, grouping arrangement, materials used, and average number of students actively engaged in the lesson (see Appendix A for further detail). The data capture teachers' use of each action within a lesson by recording the binary presence/absence of each action.

The instructional actions included in the ACOS-R are not intended to be exhaustive, given limits on the amount of detail observers can reliably notice and code within the dynamics of classroom instruction (Stodolsky, 1990). Although the instructional actions were chosen to represent the three dimensions discussed earlier (organizing, delivering literacy content, and supporting student learning), they are not organized by dimension in the coding protocol of the ACOS-R. The ACOS-R manual provided a definition and/or an explanation, along with one or more examples. For example, in the category instructional moves, one option a coder could select was "Tells/explains." In the coding manual, this action is explained as follows:

> *Tells/explains:* Telling includes explaining ideas, giving information, and providing explicit instruction. The teacher might explain a procedure or strategy (e.g., how to look up a word in a dictionary or how to summarize information in a passage).

Training Procedures

Eleven observers participated in two half-day training sessions in which they studied the coding manual, learned how to use the tablet, practiced coding video clips of literacy instruction, and discussed these with the research staff and with one another. An initial practice session involved coding short video clips of instruction in specific skill areas. The second practice session involved coding portions of a literacy block that included more than one lesson. Observers were given additional videos to practice coding with opportunities to receive feedback.

After the group training sessions, we assigned each observer to an experienced researcher/observer to carry out an observation in a classroom that was not participating in the study. The pair discussed the coding of each lesson in the literacy block to provide experience with the general instructions for carrying out observations in the classroom, opportunities to gain facility using the tablet, and time to discuss coding issues.

Inter-Observer Reliability

Two observers coded instruction of a given teacher (either independently or with two observers in the classroom at the same time) during the entire literacy block on four occasions (i.e., days equally spaced across the school year). At the beginning of the data collection period, the two observers coded instruction in one literacy block so we could determine their agreement in coding. The same procedure was used in the middle of the year to assure that the two observers assigned to each classroom remained consistent in their coding. We assessed inter-observer reliability in two ways. The first was analysis of the agreement on the partitioning of lessons and the purpose of each lesson, based on coding one literacy block. Across six pairs of observers, overall agreement was 88%. The second involved comparing agreement of coder pairs across all fields and all options within each field. Overall agreement was 87.2% with a range of 80–96% agreement.

Analytic Method

To describe teachers' instructional actions in each dimension as they relate to their students' achievement, we developed a methodological approach to accommodate three central features of our conceptual framework. First, rather than consider teachers' uses of specific actions in isolation, we drew on item response theory to

describe teachers' uses of instructional actions as guided by a set of latent dimensions (Hambleton & Swaminathan, 1985). Second, we drew on a multilevel structure to separate consistent differences among teachers in terms of their use of actions from differences among lessons within each teacher (Fox, 2010). Third, we incorporated a known mixture or multigroup component to allow for qualitative differences in teachers' uses of instruction actions across literacy skill areas. Conceptually, this feature allows us to analyze teachers' uses of actions separately for each skill area. We assembled each of these methodological approaches to form a multilevel mixture item response model with known classes (see Appendix B for the statistical model).

To address our first research question (i.e., the extent to which teachers' uses of instructional actions were less variable within than across skill areas), we compared models that ignore skill areas with those that allow for differences across skill areas. Specifically, we drew on three measures to assess the differences across skill areas: variance components, intraclass correlations coefficients, and model fit indexes. Used as a measure of stability, the intraclass correlation coefficient describes the proportions of observed variance attributable to differences among teachers and the variance attributable to differences across lessons within teachers. Higher values of intraclass correlation coefficients (i.e., closer to 1) indicate that teachers are consistent in which instructional actions they use across lessons, whereas lower values (i.e., closer to 0) indicate that the actions teachers choose to use are highly dependent on the lesson.

To address our second research question, we used the multigroup or known class mixture component of our model to describe how teachers' uses of actions differed across skill areas. As mentioned earlier, this component conceptually allows us to analyze teachers' uses of instructional actions for each skill area separately and provides a way to examine differences in teachers' uses of instructional actions across skill areas. To describe these differences, we made use of two key model parameters: action difficulty and action discrimination. In the context of teachers' uses of instructional actions, action difficulty parameters can be used to describe the proportion of lessons in which, for example, an average teacher would be expected to use a certain action in a given skill area. The analysis of action difficulties by skill area provides a way to describe the differences among skill areas by comparing the regularity with which teachers use actions in different areas.

The second feature we considered (action discrimination parameters) describe how well actions differentiate among teachers in terms of their instruction in a particular dimension. Discrimination parameters are scaled so they are positive; higher values indicate that an action can detect small differences in instruction with better precision. For example, if the value of the discrimination parameter for modeling in fluency lessons is less than that in comprehension lessons, we can presume that the action can better differentiate among teachers along the delivering literacy content dimension in comprehension lessons than in fluency lessons.

To address our third research question, we examined the relation of teachers' instruction and their students' achievement in three literacy skill areas for which we had a literacy achievement outcome. We examined the association of instruction in phonics lessons with gains on the ITBS word analysis subtest, instruction in comprehension lessons with gains on the ITBS reading comprehension subtest, and instruction in vocabulary lessons with gains on the ITBS vocabulary subtest. This provided a way to assess the extent to which teachers' uses of actions in each dimension were associated with their students' achievement gains in that area.

To assess these relationships, achievement was modeled using a hierarchical linear model (see Appendix C; Raudenbush & Bryk, 2002). At the student level, students' achievement was adjusted for grade and prior achievement in each of the three ITBS subtests. At level 2, we modeled the adjusted average achievement for each outcome as a function of the expected a posteriori estimates of teachers' stable use of actions in each dimension in each skill area derived from the aforementioned measurement model.

To address our final research question, we carried out an analysis that would make it possible to compare our skill area–specific approach to one that collapses observed instruction across literacy skill areas. That is, we constructed a measure that described teachers' instruction in each dimension, using the average number of actions a teacher used across all lessons and skill areas. Using the hierarchical linear model for achievement (see Appendix C), we re-estimated the correlation between each of the instructional dimensions and student achievement subtests and compared them with those of the skill area–specific indexes.

Results

Variance in Teachers' Instruction

The first research question focused on the variability with which teachers use actions within and across skill areas. To address this question, we first partitioned the observed variation in instruction into lesson and teacher components but without taking the skill area of each lesson into account. The results are shown in Table 2. The first row describes the variance attributable to differences among lessons within teachers, whereas the second row

TABLE 2
Variability of Instruction by Instructional Dimension

Components of variability	Delivering literacy content		Supporting student learning		Organizing	
	Collapsed	Skill area–specific	Collapsed	Skill area–specific	Collapsed	Skill area–specific
Lesson-to-lesson variance	1.83	1.56	0.45	0.15	3.25	2.40
Intraclass correlation coefficient	0.35	0.39	0.69	0.87	0.24	0.29
Proportion of lesson-to-lesson variance explained	0.15		0.67		0.26	

Note. Collapsed refers to the model in Appendix B with parameters constrained to be equal across skill areas. *Skill area–specific* refers to the full model in Appendix B. Teacher-level variance was set to 1 to identify the model. The intraclass correlation coefficient describes the stability of teachers' choices by expressing the proportion of variance in teachers' instruction attributable to consistent differences among teachers in terms of their uses of instructional actions.

describes the proportion of the total variability attributable to stable differences among teachers (i.e., intraclass correlation coefficient). For each of the three dimensions, results indicated substantial lesson-to-lesson variability within teachers (as shown in the Collapsed columns for each dimension in the table).

As described by the intraclass correlation coefficient, teachers' uses of instructional actions in the organizing dimension were the most variable. Only 24% of the observed variation was attributable to stable differences among teachers' choices, while the remaining 76% was attributable to differences among lessons within teachers. Approximately 35% of the observed variation in the delivering literacy content dimension was attributable to consistent differences among teachers, whereas 69% of the variation in actions to support student learning was attributable to consistent differences among teachers. These results demonstrate just how large lesson-to-lesson variation is when skill area is not taken into account.

We then considered the extent to which teachers' instruction was less variable across lessons within a skill area than across all lessons (without regard for skill areas). The results supported our expectation that there would be less variability when we examined data by skill area than across skill areas. This is apparent in Table 2 from the decrease in lesson-to-lesson variance (and increase in intraclass correlation coefficients) when moving from the Collapsed columns to the Skill Area–Specific columns.

The third row in this table summarizes the decreased variability by describing the proportion of lesson-to-lesson variance explained by systematic difference among skill areas (i.e., comparing the within-teacher variation for the Collapsed and Skill Area–Specific columns). Differences in teachers' uses of instructional actions by skill area explained approximately 67% of the lesson-to-lesson variation within teachers in the supporting student learning dimension, 26% of the lesson-to-lesson variation in the organizing dimension, and 15% of the lesson-to-lesson variation in the delivering literacy content dimension. Furthermore, formal comparisons of model fit indexes provided strong additional support that teachers' uses of the instructional actions were significantly more consistent across lessons within the same skill area than across lessons in all literacy skill areas (as shown by the statistically significant differences in deviances in Table 3).

TABLE 3
Comparison of Skill Area–Specific and Collapsed Models by Dimension

Model	Deviance	Difference between deviances	Degrees of freedom
Delivering literacy content			
Skill area–specific model	12,274	202*	20
Collapsed across skill areas model	12,476		
Supporting student learning			
Skill area–specific model	8,514	109*	16
Collapsed across skill areas model	8,623		
Organizing			
Skill area–specific model	10,004	57*	20
Collapsed across skill areas model	10,061		

Note. Skill area–specific model refers to the full model in Appendix B. *Collapsed across skill areas model* refers to the model in Appendix B with parameters constrained to be equal across literacy skill areas.
*$p < .05$ using the likelihood ratio test.

Differences in Teachers' Uses of Instructional Actions

Having found evidence that teachers' uses of instructional actions in each dimension were less variable within than across skill areas, we investigated two ways in which teachers' uses of the actions differed by skill area. First, we examined the regularity with which an average teacher would be expected to use each of the actions in each skill area. Table 4 shows the breakdown of the proportion of lessons average teachers are expected to use the instructional actions by skill area. Overall, teachers tended to use actions in the delivering literacy content dimension in most lessons except for those focused on fluency. For example, an average teacher would be expected to use the action "asking questions for evaluation" in about 88% of comprehension lessons, 77% of vocabulary lessons, 77% of phonics lessons, and 73% of writing lessons, but only 34% of fluency lessons. Our results also suggested that an average teacher would not use actions in the supporting student learning and organizing dimensions in most lessons, although this again differed by skill area. Finally, another noteworthy finding was that teachers used a wide variety of instructional actions in teaching reading comprehension.

Although descriptions that collapse across teachers but not skill areas (e.g., results presented in Table 4) show some variation in teachers' use of actions across skill areas, such descriptions may still undervalue the magnitude of variation in individual teachers' uses of these actions across skill areas. To illustrate this problem, we created a graphic display (see Figure 1) of within-teacher variation for a random selection of five teachers (labeled A–E) by plotting the average number of actions each teacher used across lessons by skill area. An important finding is that teachers' use of instructional actions was highly variable across skill areas. For example, whereas Teacher A regularly used about four delivering literacy content actions in vocabulary lessons, the teacher regularly used only about one of the four actions from this dimension in fluency lessons. This pattern of variation in use of instructional actions by skill area is evident for all five teachers. Together, Table 4 and Figure 1 illustrate the concerns that motivated our study, using our observational data.

Second, the results suggested that the discrimination parameters (ability of actions to differentiate among teachers) varied by skill area for each dimension, as shown in Table 5. Although actions in the delivering literacy content dimension were best able to capture differences among teachers in phonics lessons, they poorly described differences in fluency lessons. Similarly, actions in the supporting student learning dimension best differentiated among teachers in phonics lessons but did not effectively describe teachers' instruction in vocabulary

TABLE 4
Proportion of Lessons an Average Teacher Would Be Expected to Use Each Instructional Action by Skill Area

Instructional action by dimension	Phonics	Fluency	Writing	Comprehension	Vocabulary
Delivering literacy content					
Telling/explaining	.77	.38	.78	.82	.75
Modeling/coaching	.65	.44	.66	.67	.59
Asking questions for evaluation	.77	.34	.73	.88	.77
Providing practice or review for activities	.91	.51	.81	.75	.75
Supporting student learning					
Fostering discussion	.05	.12	.18	.40	.40
Assessing students' work; providing feedback	.22	.22	.33	.33	.41
Giving students an opportunity to ask questions	.02	.05	.14	.19	.32
Organizing					
Explaining the purpose of the lesson	.20	.28	.33	.38	.23
Explaining the value of the lesson	.09	.26	.19	.10	.05
Giving directions for an activity	.56	.44	.65	.78	.77
Providing a wrap-up or summary	.07	.23	.17	.12	.04

FIGURE 1
Within-Teacher Variation for Five Teachers (A–E) by Skill Area

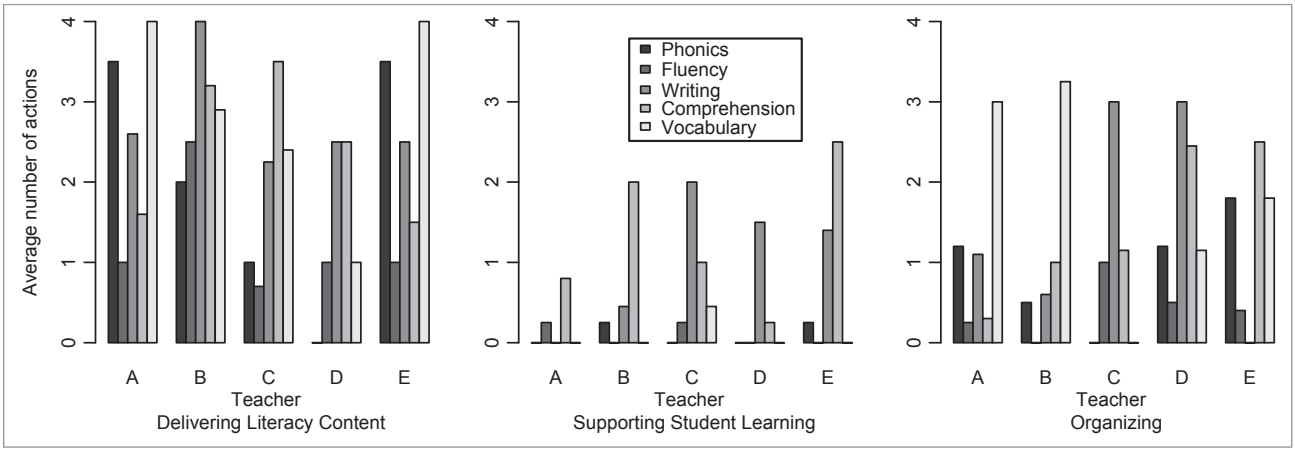

TABLE 5
Discrimination Parameter Estimates by Dimension and Skill Area

Dimension	Phonics	Fluency	Writing	Comprehension	Vocabulary
Delivering literacy content	1.05	0.47	0.82	0.80	0.77
Supporting student learning	0.91	0.56	0.60	0.67	0.23
Organizing	0.61	0.24	0.46	0.78	0.85

lessons. In contrast, we found that actions in the organizing dimension were best suited to identify differences among teachers in vocabulary and comprehension lessons but poorly suited to identify differences in fluency lessons.

Relation of Dimensions and Students' Achievement

Our third research question focused on the extent to which teachers' uses of actions in three of the literacy skill areas were associated with students' achievement in that skill area. The results (standardized regression coefficients and standard errors) are shown in Table 6. For lessons in phonics, teachers' uses of actions in each of the dimensions accounted for significant gains on the ITBS subtest for word analysis. For lessons in reading comprehension, instructional actions in the delivering literacy content and supporting student learning dimensions accounted for significant gains on the ITBS reading comprehension subtest. For lessons in vocabulary, teachers' uses of instructional actions in the delivering literacy content and supporting student learning dimensions were associated with gains on the ITBS vocabulary subtest and were similar in magnitude to the associations found in comprehensions lessons.

Comparison With Conventional Approach

Our final question was whether conventional methods would provide similar findings regarding the relationship of teachers' uses of actions and their students' achievement. To answer this question, we reestimated the association between student achievement on each subtest and indexes of teachers' uses of actions in each dimension, using indexes of teachers' instruction collapsed across literacy skill areas. The results, presented in Table 7, show that for most literacy outcomes and dimensions, teachers' uses of instructional actions (collapsed across skill areas) were not significantly related to their students' achievement.

By comparison, the indexes that describe teachers' uses of instructional actions in specific skill areas (see Table 6) showed a much stronger association with their students' gains in the three skill areas. For example, examination of teachers' uses of delivering literacy content in phonics lessons, skill area–specific indexes (see Table 6) indicated that there was a significant relationship with an effect size of about .08, whereas the conventional method suggested that this relationship was not statistically different from 0 (see Table 7).

TABLE 6
Standardized Regression Coefficients Describing the Relation Between Teachers' Skill Area–Specific Use of Instructional Actions in Each Dimension and Their Students' Achievement in That Area

	Coefficient (and standard error)		
Variable	Phonics[a]	Comprehension[b]	Vocabulary[c]
Iowa Test of Basic Skills word analysis pretest	.32* (.02)	.16* (.02)	.15* (.02)
Iowa Test of Basic Skills comprehension pretest	.32* (.02)	.44* (.02)	.39* (.02)
Iowa Test of Basic Skills vocabulary pretest	.15* (.02)	.20* (.02)	.23* (.02)
Grade 3	−.04 (.05)	.03 (.04)	.32* (.05)
Delivering literacy content	.08* (.02)	.07* (.02)	.09* (.02)
Supporting student learning	.06* (.03)	.05* (.02)	.03^ (.02)
Organizing	.12* (.03)	.01 (.02)	−.02 (.02)

[a]Outcome used was the Iowa Test of Basic Skills word analysis subtest. [b]Outcome used was the Iowa Test of Basic Skills comprehension subtest. [c]Outcome used was the Iowa Test of Basic Skills vocabulary subtest.
*$p < .05$. ^$p < .10$.

TABLE 7
Standardized Regression Coefficients Describing the Relation Between the Average Number of Actions Teachers Use in a Lesson for Each Dimension and Students' Achievement

	Coefficient (and standard error)		
Variable	Phonics[a]	Comprehension[b]	Vocabulary[c]
Iowa Test of Basic Skills word analysis pretest	.32* (.02)	.17* (.02)	.15* (.02)
Iowa Test of Basic Skills comprehension pretest	.31 (.02)	.44* (.02)	.39* (.02)
Iowa Test of Basic Skills vocabulary pretest	.16 (.02)	.20* (.02)	.24* (.02)
Grade 3	.06 (.05)	.04 (.05)	.33* (.05)
Delivering literacy content	.01 (.03)	.05* (.02)	.06^ (.03)
Supporting student learning	.03 (.03)	.03 (.02)	.00 (.03)
Organizing	.08* (.03)	−.02 (.02)	−.03 (.03)

[a]Outcome used was the Iowa Test of Basic Skills word analysis subtest. [b]Outcome used was the Iowa Test of Basic Skills comprehension subtest. [c]Outcome used was the Iowa Test of Basic Skills vocabulary subtest.
*$p < .05$. ^$p < .10$.

Discussion

The purpose of this study was to investigate a conceptual and methodological approach that would help address current problems in deriving information about effective literacy instruction from observational studies. The approach is based on the premise that teachers' instructional actions are more reliable and valid within than across literacy skill areas. Our expectation was that teachers' uses of instructional actions would be more consistent within than across skill areas; we also anticipated that observed patterns of instruction within a skill area would be more significantly related to students' literacy achievement in that skill area than would descriptions of instruction that were collapsed across skill areas.

The results, overall, supported these expectations. Thus, we have evidence that our proposed approach to measuring and analyzing early literacy instruction could constitute a significant methodological advance, one that holds the promise of improving research efforts to identify effective early literacy instruction in different skill areas. In what follows, we explore the results and their implications for future observational studies of early literacy.

Dimensions of Literacy Instruction Within Skill Areas

One important feature of our approach to studying observed instruction was the use of key dimensions of effective literacy instruction that might differentially

characterize teachers' instruction in different literacy skill areas. The results indicated systematic differences by literacy skill area for the three dimensions. For actions in the delivering literacy content dimension, the results showed that teachers used actions regularly in all areas except fluency. The frequent use of actions teacher take in delivering literacy content reflects research findings that suggest the importance of explicit, guided instruction in early elementary literacy (e.g., Foorman & Torgesen, 2001; Roehler & Duffy, 1991; Taylor et al., 2003).

The relative infrequency of teacher-led instruction in fluency is striking but might not be surprising. Although researchers have identified a number of effective teacher-led approaches to fluency instruction, Rasinski, Reutzel, Chard, and Linan-Thompson (2011) reported that for many years, teachers sought to improve students' fluency simply by giving them time for independent reading practice. Given our findings, we might infer that teachers in our study placed particular emphasis on practice as a way to develop fluency.

Teachers' use of supporting student learning actions was uncommon in lessons in all literacy skill areas. This finding runs counter to the view that elementary students need guidance, monitoring, and involvement to be appropriately attentive to their work (e.g., Roehler & Duffy, 1991). Perhaps the limited attention to students' engagement reflects a view of literacy instruction promoted by the Reading First program. Still, it is important to note that although actions in this dimension were uncommon, they nonetheless contributed to students' gains in phonics, comprehension, and vocabulary achievement. That is, even modest use of instructional actions in the supporting student learning dimension contributed to gains in students' literacy achievement. As Guthrie and his colleagues (2006) found, students' motivation and engagement mediate their literacy acquisition and contribute to their progress and self-regulation in reading and writing.

With regard to actions in the organizing dimension, the results indicated that teachers regularly gave directions but seldom explained the purpose or value of a lesson or provided a wrap-up of the lesson. There was some variation by skill areas. To some extent, we found greater attention to actions in the organizing dimension in reading comprehension than in fluency or phonics. This finding might provide support for Rosenshine's (1995) distinction of instruction in more and less structured domains. Organization would seem to be critical in all areas of literacy as it contributes to students' understanding of what they are learning (or are about to learn) and why. Previous studies have supported this expectation (e.g., Cameron et al., 2005; Guthrie, 2004). For example, Duffy and his colleagues (1986) found that students' understanding of lesson content depended on the organizing instructions given by their teacher.

The results are compatible with descriptions of effective instruction in specific skill areas. For example, researchers tend to stress the importance of practice and feedback in effective teaching of phonics and word reading (e.g., Brady, 2011; Foorman & Connor, 2011; Foorman & Torgesen, 2001), whereas modeling comprehension strategies and fostering discussion play an important role in teaching reading comprehension effectively (e.g., Shanahan et al., 2010). Such descriptive results might help us envision how analysis of instruction within skill areas could be used to distinguish more and less effective literacy instruction and could contribute to teachers' understanding of literacy instruction in specific literacy skill areas.

Although our findings are compatible with theory and research in early literacy, it is important to remember that the purpose of our study was not to provide a detailed description of effective early literacy instruction but to test aspects of a new approach to gathering and analyzing data from classroom observations. The characterization of instruction by skill area that comes from this study appears to provide sufficient evidence of the validity of the approach to warrant further study. However, as pointed out earlier, other researchers may choose to include additional dimensions or instructional actions. They would also want to take characteristics of the students in each classroom into account to examine the extent to which teachers' instruction is sensitive to the needs of the students (e.g., Connor et al., 2009).

Association of Instruction and Achievement by Skill Area

To examine the predictive validity of our proposed approach to gathering and analyzing early reading instruction, we examined the extent to which teachers' engagement in the three dimensions in a specific skill area was associated with students' achievement gains in the areas of vocabulary, phonics, and reading comprehension. Although the magnitude of the relationships between teachers' actions in each dimension and achievement in a specific skill area was small, the cumulative impact on achievement was .26 standard deviation in phonics, .12 in comprehension, and .12 in vocabulary (see Table 6).[1]

Benchmarking these impacts against the average reported achievement gains in these grades, we found that in phonics, a .26 standard deviation gain is roughly equal to almost eight weeks of extra growth (Hoover et al., 2003). Similarly, the effects in comprehension and vocabulary are associated with about four extra weeks of growth. Because our sample focused on high-poverty districts where low achievement tends to be a persistent problem, such gains are particularly meaningful in that

they might help narrow the achievement gap associated with economically challenged schools.

To determine the extent to which our approach constitutes an improvement in the design and analysis of observational studies, we compared the gain in students' achievement when teachers' instruction was or was not collapsed across literacy skill areas. Comparing the results (see Tables 6 and 7), we can see that descriptions of teachers' uses of instructional actions in specific skill areas were more strongly associated with student achievement in that skill area than were general descriptions of instruction that ignored differences among literacy skill areas. For example, descriptions of teachers' actions in the supporting student learning dimension that were collapsed across literacy skill areas found no relationship between teachers' uses and their students' achievement in phonics. In contrast, descriptions of teachers' actions in the supporting student learning dimension that focused specifically on teachers' uses of these actions in phonics lessons found a statistically significant relationship.

Exchangeability of Instruction Across Skill Areas

Although theory would suggest that teachers vary across skill areas in the instructional actions they use to deliver content to students, researchers have commonly collapsed observation data across skill areas to conduct analyses (e.g., Taylor et al., 2003). As discussed earlier, the soundness of this approach rests on the extent to which teachers' uses of actions across skill areas are interchangeable. Our findings suggest that literacy skill areas may not be exchangeable.

Evidence to support our approach comes from differences in teachers' uses of instructional actions in literacy skill areas, stronger relationships between instruction in three skill areas and achievement gains in those skill areas, and the variability in the effectiveness of teachers' uses of actions across skill areas. Put differently, the results suggest that descriptions of instruction that collapse across skill areas potentially underrepresent important differences in teachers' uses of specific actions in specific literacy skill areas.

The results also raise questions about the validity of collapsing assessments of teaching across lessons in different subjects. Recent investigations into teacher quality have routinely collapsed quality ratings across subjects such as mathematics and English language arts (MET Project, 2012, 2013). Collapsing teacher quality indexes across lessons in different subject areas inherently assumes that a teacher's quality is exchangeable (i.e., has the same distribution) across lessons in different subject areas. However, it is quite possible that a teacher's quality varies by subject area (i.e., lessons are not interchangeable across subject areas). For instance, using the Classroom Assessment Scoring System (Pianta, La Paro, & Hamre, 2008), teachers may differ in the quality with which they offer an instructionally supportive environment between mathematics and English language arts lessons because their content knowledge and training differ by subject. Given evidence that instruction varies by literacy skill area, we believe further research into the exchangeability of teacher quality across subjects is warranted.

Study Design, Limitations, and Recommendations for Further Research

Although our findings suggest that we have identified promising ways to improve what we learn about effective literacy instruction from classroom observations, there are important limitations to the design of the study. First, we would have liked to take into account the characteristics of students in each lesson, but the constraints of the sample size and data collection methods made this impossible. The studies carried out by Connor and her colleagues (2004, 2009) offer a model for others who would like to determine whether instructional choices were appropriate for the particular students. In future studies, other factors might also be taken into account, such as the quality with which teachers used actions, the frequency of disruptions, and the appropriateness of reading materials.

Second, the study focused just on grades 2 and 3 and a relatively small number of teachers. Further research is needed to explore the value of the approach to studying literacy instruction in different grades and schools. Third, as with all observational studies, what you look for is what you learn about. Thus, our results apply specifically to the three dimensions and instructional actions that represent these. As noted earlier, depending on the purpose of a study, researchers developing an observation system might choose to include different or additional dimensions and instructional actions.

The study we report herein represents an effort to investigate an approach to designing and analyzing observation studies that can provide trustworthy and meaningful information about effective early literacy instruction. Given the increasingly frequent use of classroom observations as a way to evaluate teachers' teaching, it is important that teachers and researchers understand both the promise and the challenges of carrying out observational studies (Hoffman et al., 2011; MET Project, 2012, 2013). Although the results of this study suggest that our approach may represent a methodological breakthrough, further study is needed to refine and examine theoretically and empirically sound methods to measure effective literacy instruction.

NOTES

This study was made possible by a Teacher Quality grant from the Institute for Education Sciences (IES; award R305M050087); however, IES is not responsible for the design and execution of the study, the interpretation, or the results. In addition, we are grateful for the support of Michigan's Reading First program directors, and we would like to thank the participating elementary teachers for welcoming us into their classrooms.

[1] From Table 6, the contribution of delivering literacy content, supporting student learning, and organizing to word analysis achievement was .08, .06, and .12, respectively; the contribution of the delivering literacy content and supporting student learning dimensions to reading comprehension achievement was .07 and .05, respectively; and the contribution of the delivering literacy content and supporting student learning dimensions to vocabulary achievement was .09 and .03, respectively.

REFERENCES

Barr, R., & Dreeben, R. (with Wiratchai, N.). (1983). *How schools work*. Chicago: University of Chicago Press.

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Brady, S.A. (2011). Efficacy of phonics teaching for reading outcomes: Indications from post-NRP research. In S.A. Brady, D. Braze, & C.A. Fowler (Eds.), *Explaining individual differences in reading: Theory and evidence* (pp. 69–96). New York: Psychology.

Brophy, J., & Good, T.L. (1986). Teacher behavior and student achievement. In M.C. Whittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Simon & Schuster.

Cameron, C.E., Connor, C.M., & Morrison, F.J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology*, *43*(1), 61–85. doi:10.1016/j.jsp.2004.12.002

Cameron, C.E., & Morrison, F.J. (2011). Teacher activity orienting predicts preschoolers' academic and self-regulatory skills. *Early Education and Development*, *22*(4), 620–648. doi:10.1080/10409280903544405

Carlisle, J.F., Kelcey, B., Berebitsky, D., & Phelps, G. (2011). Embracing the complexity of reading instruction: A study of the effects of teachers' instruction on students' reading comprehension. *Scientific Studies of Reading*, *15*(5), 409–439. doi:10.1080/10888438.2010.497521

Connor, C.M., Morrison, F.J., Fishman, B.J., Ponitz, C.C., Glasney, S., Underwood, P.S., et al. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher*, *38*(2), 85–99. doi:10.3102/0013189X09332373

Connor, C.M., Morrison, F.J., & Petrella, J.N. (2004). Effective reading comprehension instruction: Examining child × instruction interactions. *Journal of Educational Psychology*, *96*(4), 682–698. doi:10.1037/0022-0663.96.4.682

Correnti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal*, *44*(2), 298–339. doi:10.3102/0002831207302501

Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational Researcher*, *38*(7), 518–521. doi:10.3102/0013189X09350881

Duffy, G.G., Roehler, L.R., & Rackliffe, G. (1986). How teachers' instructional talk influences students' understanding of lesson content. *The Elementary School Journal*, *87*(1), 3–16. doi:10.1086/461476

Foorman, B.R., & Connor, C.M. (2011). Primary grade reading. In M.L. Kamil, P.D. Pearson, E.B. Moje, & P.P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 136–156). New York: Routledge.

Foorman, B.R., Schatschneider, C., Eakin, M.N., Fletcher, J.M., Moats, L.C., & Francis, D.J. (2006). The impact of instructional practices in grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology*, *31*(1), 1–29. doi:10.1016/j.cedpsych.2004.11.003

Foorman, B.R., & Torgesen, J. (2001). Critical elements of classroom and small-group instruction promote reading success in all children. *Learning Disabilities Research & Practice*, *16*(4), 203–212. doi:10.1111/0938-8982.00020

Fox, J.P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer. doi:10.1007/978-1-4419-0742-4

Good, T.L., & Mulryan, C. (1990). Teacher ratings: A call for teacher control and self-evaluation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 191–215). Newbury Park, CA: Sage.

Guthrie, J.T. (2004). Teaching for literacy engagement. *Journal of Literacy Research*, *36*(1), 1–30. doi:10.1207/s15548430jlr3601_2

Guthrie, J.T., Wigfield, A., Humenick, N.M., Perencevich, K.C., Taboada, A., & Barbosa, P. (2006). Influences of stimulating tasks on reading motivation and comprehension. *The Journal of Educational Research*, *99*(4), 232–246. doi:10.3200/JOER.99.4.232-246

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hill, H.C., Charalambous, C., & Kraft, M.A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. doi:10.3102/0013189X12437203

Hoffman, J.V. (1991). Teacher and school effects in learning to read. In R. Barr, M.L. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 911–950). New York: Longman.

Hoffman, J.V., Maloch, B., & Sailors, M. (2011). Researching the teaching of reading through direct observation. In M.L. Kamil, P.D. Pearson, E.B. Moje, & P.P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 3–33). New York: Routledge.

Hoffman, J.V., Sailors, M., Duffy, G.R., & Beretvas, S.N. (2004). The effective elementary classroom literacy environment: Examining the validity of the TEX-IN3 observation system. *Journal of Literacy Research*, *36*(3), 303–334. doi:10.1207/s15548430jlr3603_3

Hoover, H.D., Dunbar, S.B., Frisbee, D.A., Oberly, K.R., Ordman, V.L., Naylor, R.J., et al. (2003). *Iowa Test of Basic Skills: Guide to research and development*. Ithaca, IL: Riverside.

Klein, S.P., & Stecher, B.M. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, *11*(2), 121–137. doi:10.1207/s15324818ame1102_1

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

MET Project. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved April 26, 2013, from www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

MET Project. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved April 26, 2013, from www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Pianta, R.C., & Hamre, B.K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. doi:10.3102/0013189X09332374

Pianta, R.C., La Paro, K.M., & Hamre, B.K. (2008). *Classroom Assessment Scoring System manual, pre-K*. Baltimore: Brookes.
Piasta, S.B., Connor, C.M., Fishman, B.J., & Morrison, F.J. (2009). Teachers' knowledge of literacy concepts, classroom practices, and student reading growth. *Scientific Studies of Reading*, *13*(3), 224–248. doi:10.1080/10888430902851364
Porter, A.C., & Brophy, J. (1988). Synthesis of research on good teaching: Insights from the work of the Institute for Research on Teaching. *Educational Leadership*, *45*(8), 74–85.
Pressley, M., Wharton-McDonald, R., Raphael, L.M., Bogner, K., & Roehrig, A. (2002). Exemplary first-grade teaching. In B.M. Taylor & P.D. Pearson (Eds.), *Teaching reading: Effective schools, accomplished teachers* (pp. 73–88). Mahwah, NJ: Erlbaum.
Rasinski, T.V., Reutzel, D.R., Chard, D., & Linan-Thompson, S. (2011). Reading fluency. In M.L. Kamil, P.D. Pearson, E.B. Moje, & P.P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 286–319). New York: Routledge.
Rasinski, T., Homan, S., & Biggs, M. (2009). Teaching reading fluency to struggling readers: Method, materials, and evidence. *Reading & Writing Quarterly*, *25*(2/3), 192–204. doi:10.1080/10573560802683622
Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
Roehler, L.R., & Duffy, G.G. (1991). Teachers' instructional actions. In R. Barr, M.L. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 861–910). New York: Longman.
Rosenshine, B. (1995). Advances in research on instruction. *The Journal of Educational Research*, *88*(5), 262–268. doi:10.1080/00220671.1995.9941309
Rosenshine, B., & Stevens, R. (1984). Classroom instruction in reading. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 745–798). New York: Longman.
Seidel, T., & Shavelson, R.J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499. doi:10.3102/0034654307310317
Shanahan, T., Callison, K., Carriere, C., Duke, N.K., Pearson, P.D., Schatschneider, C., et al. (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide* (NCEE 2010-4038). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
Shavelson, R.J., Webb, N.M., & Burstein, L. (1986). Measurement of teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50–91). New York: Simon & Schuster.
Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23.
Snow, C.E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.
Snow, C.E., Burns, M.S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
Stodolsky, S.S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175–190). Newbury Park, CA: Sage.
Taylor, B.M., Pearson, P.D., Peterson, D.S., & Rodriguez, M.C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal*, *104*(1), 3–28. doi:10.1086/499740
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*(1), 40–69. doi:10.1598/RRQ.40.1.3
U.S. Department of Education. (2002). *Guidance for the Reading First program*. Washington, DC: Office of Elementary and Secondary Education, U.S. Department of Education. Retrieved April 26, 2013, from www2.ed.gov/programs/readingfirst/guidance.pdf
Webb, N.M., Shavelson, R.J., & Haertel, E.H. (2006). 4 reliability coefficients and generalizability theory. In C.R. Rao (Vol. Ed.) & S. Sinharay (Vol. & Gen. Ed.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 81–124). Amsterdam: North-Holland. doi:10.1016/S0169-7161(06)26004-8

Submitted July 9, 2012
Final revision received April 8, 2013
Accepted April 11, 2013

BEN KELCEY (corresponding author) is an assistant professor in the College of Education, Criminal Justice, and Human Services at the University of Cincinnati, Ohio, USA; e-mail ben.kelcey@gmail.com. His research interests include the development of measurement and quantitative research methods to understand effective teaching and teachers.

JOANNE F. CARLISLE is a professor emerita in the School of Education at the University of Michigan, Ann Arbor, USA; e-mail jfcarl@umich.edu. Her research interests focus on the relation of language and literacy development with a special interest in children for whom language and literacy acquisition presents unusual challenges.

APPENDIX A

Coding Categories and Options in the ACOS-R

Purpose

Options: Phonological awareness; phonics, word reading; fluency; writing; reading comprehension; vocabulary; assessment; centers; other

Grouping

Options: Teacher and students working as a whole class; teacher working with whole class, students working in small groups; teacher working with whole class,

students working individually; teacher working with a small group, students working in small groups or individually; teacher working with an individual student, students working in small groups or individually; other

Word Meaning

Options: The teacher defines a word or word part; the teacher states or reads a sentence containing the word; the teacher asks students to explain a word's meaning; the teacher asks students to use a word in a sentence; the teacher fosters discussion of a word's meaning.

Materials

Options: Anthology; trade book or leveled reader; writing materials; manipulatives; computer, projector, or other technology; chalkboard, interactive whiteboard, pads of paper, or pocket chart; other; none

Instructional Moves

Options: Tells, explains; models, coaches; asks questions for evaluation; fosters or initiates discussion; provides practice or review activities; assesses student learning; explains purpose of lesson; explains value or relevance of lesson; gives directions for activity; gives students an opportunity to ask questions; provides a wrap-up or summary

Engagement

A student is engaged if he or she is participating in any literacy activity suitable for the lesson or directed activity indicated by the teacher.

Options: High (90% of the students are on task); medium (70–90% of the students are on task); low (less than 70% of the students are on task)

NOTE

An additional source of information was explanations or descriptions of activities and lessons that observers entered in text boxes.

APPENDIX B

Multigroup Multilevel Item Response Measurement Model for Instruction

We can express the model as

$$P(Y^{d}_{ijkg} = 1|\theta) = \frac{\exp[a^{d}_{g}(\theta^{d}_{kg} + \theta^{d}_{jkg} - b^{d}_{ig})]}{1 + \exp[a^{d}_{g}(\theta^{d}_{kg} + \theta^{d}_{jkg} - b^{d}_{ig})]} \qquad (1)$$

where $P(Y^{d}_{ijkg} = 1 \mid \theta)$ is the conditional probability that teacher k used instructional action i in dimension d (teacher-directed instruction, support for student learning, or pedagogical structure) in lesson j with skill area focus g (known class). θ^{d}_{kg} is teacher k's stable use of actions in dimension d in literacy skill area g with matching discrimination parameter a^{d}_{g}, which describes the strength with which the measured actions describe dimension d for skill area g. θ^{d}_{jkg} is the extent to which lesson j in literacy skill area g for teacher k deviates from teacher k's stable use of actions in that dimension and skill area. Finally, b^{d}_{ig} is the difficulty or expected regularity with which an average teacher might employ action i in dimension d in skill area g.

To identify the model, the dimensions were specified to have a multivariate normal distribution with the scale of each teacher level dimension in each literacy skill area fixed to a mean of 0 with unit variance, and each lesson level dimension was centered at 0 with lesson-level variances assumed to be equal across skill areas.

APPENDIX C

Hierarchical Linear Model for Student Achievement

At the student level, students' achievement was adjusted for grade and prior achievement in each of the ITBS subtests so that

$$Y_{ij}^{(g)} = \pi_{0j}^{(g)} + \sum_{p=1}^{n=4} \pi_{p}^{(g)} X_{pij} + \varepsilon_{ij}^{(g)} \quad (2)$$

where $Y_{ij}^{(g)}$ is the ITBS posttest score for student i in classroom j for literacy skill area g; $\pi_{0j}^{(g)}$ is the average student score adjusted for grade; the prior achievement variables, X and $\pi_{p}^{(g)}$, are the corresponding coefficients for the prior achievement variables; and $\varepsilon_{ij}^{(g)}$ has a normal distribution with mean 0 and variance $\sigma_{(g)}^{2}$.

At level 2, we modeled the adjusted average achievement for outcome g, $\pi_{0j}^{(g)}$, as a function of the expected a posteriori estimates of teachers' stable use of actions in each dimension in each skill area derived from the measurement model in equation 1 (Bartholomew & Knott, 1999):

$$\pi_{0j}^{(g)} = \beta_{00}^{(g)} + \beta_{01}^{(g)} DLC_{j}^{(g)} + \beta_{02}^{(g)} SSL_{j}^{(g)} + \beta_{03}^{(g)} O_{j}^{(g)} + r_{0j}^{(g)} \quad (3)$$

where $\beta_{00}^{(g)}$ is the average adjusted achievement level for skill area g, and $\beta_{01}^{(g)}$, $\beta_{02}^{(g)}$, and $\beta_{03}^{(g)}$ are associations between teachers' use of delivering literacy content (*DLC*), supporting student learning (*SSL*), and organizing (*O*) in lessons in literacy skill area g and achievement on the outcome for that skill area. Finally, $r_{0j}^{(g)}$ is the normally distributed random effect of teacher j with mean 0 and variance τ^{2}.