Early Childhood Research Quarterly 27 (2012) 316–328

Contents lists available at SciVerse ScienceDirect

Early Childhood Research Quarterly

Reliability and validity of the Classroom Observations of Student–Teacher Interactions (COSTI) for kindergarten reading instruction

Keith Smolkowski*, Barbara Gunn

Oregon Research Institute, United States

ARTICLE INFO

Article history:
Received 25 January 2010
Received in revised form 1 July 2011
Accepted 10 September 2011

Keywords:
Observations
Reliability
Validity
Reading instruction
Instructional quality
Student independent practice

ABSTRACT

This paper describes the technical adequacy and potential uses of an observation system used to measure the quality of literacy instruction in kindergarten classrooms. The Classroom Observations of Student–Teacher Interactions (COSTI) documents the frequency of four student–teacher interactions during beginning reading instruction: explicit teacher demonstrations, student independent practice, student errors, and teacher corrective feedback. Data were collected during kindergarten reading instruction, and the analyses address reliability, stability of the coded teaching behaviors, and predictive validity. Results indicated that data could be collected reliably and that teachers' provision of opportunities for independent student practice was stable across the school year. Student independent practice opportunities also predicted gains in several important reading outcomes. Implications are discussed, including potential uses of the instrument for providing teachers with feedback on their literacy instruction and for extending the knowledge base on effective literacy instructional practices.

1. Introduction

Teaching and learning basic skills, such as beginning reading, require a high level of interaction between teachers and their students. Interactions such as demonstrating new skills, affording opportunities for independent practice, recognizing mistakes, and providing corrective feedback serve as an essential vehicle for teaching children fundamental concepts and skills (Archer & Hughes, 2011). In order to provide teachers with feedback on their instruction in basic skills and to document student–teacher interactions for research on effective instructional practices, it is important to reliably measure student–teacher interactions during instruction and demonstrate that teachers' ability to provide sufficient and timely interactions contributes to student learning. The Classroom Observations of Student–Teacher Interactions (COSTI) instrument was developed to quantify the rates of specific instructional interactions that occur between teachers and their students. This paper describes the COSTI and reports the reliability and validity of the instrument from a study on beginning reading instruction.

1.1. Role of instructional interactions in beginning reading instruction

Instructional interactions between students and teachers are thought to lay the foundation for acquiring initial reading skills based on basic principles of learning and retention (Carver & Klahr, 2001; Ebbinghaus, Ruger, & Bussenius, 1913). Explicit teacher demonstration of new skills and frequent opportunities for student independent practice, coupled with specific, corrective feedback on student errors are particularly important during beginning reading instruction because they provide children with a basic foundation for acquisition of early reading skills (Aylward et al., 2003; Goswami, 2004; Shaywitz, Morris, & Shaywitz, 2008; Simos et al., 2002; Stevens, Fanning, Coch, Sanders, & Neville, 2008; Temple et al., 2003). Teacher demonstrations and corrective feedback make clear to students what they are learning and what it looks and sounds like when accomplished correctly. Independent practice helps students gain mastery and fluency with newly learned skills, concepts, and vocabulary. Evidence from this body of research suggests that these interactions, particularly practice, appear to be particularly important (Ericsson, Roring, & Nandagopal, 2007; Fields, 2005; Meltzoff, Kulh, Movellan, & Sejnowski, 2009; Sutherland, Alder, & Gunter, 2003; Swanson & O'Connor, 2009). Findings from these distinct but related bodies of research suggest that the rates of specific student–teacher interactions are related to the acquisition of basic academic skills such as decoding, math facts, and spelling. These interactions are particularly important during preschool and grades K-2 when children

* Corresponding author at: Oregon Research Institute, 1715 Franklin Boulevard, Eugene, OR 97403, United States. Tel.: +1 541 484 2123; fax: +1 541 484 1108.
E-mail address: keiths@ori.org (K. Smolkowski).

0885-2006/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ecresq.2011.09.004

are taught the basic skills that will be essential in order for them to learn and apply more advanced knowledge and skills throughout their school and working careers. In this section, we define the interactions and describe the theoretical and research support for their link to student learning.

1.1.1. Teacher demonstration

Teacher demonstration is a key feature of explicit instruction (Archer & Hughes, 2011). Teacher demonstration is defined as the teacher giving students new information or showing students how to apply a new skill, such as “the letter *m* makes the sound/mmm/” or “I’m going to sound out the word *man*,/mmm//aaa//nnn/.” Teacher demonstration is used when teaching a new skill, when students need more practice, or when students do not reply or make an error. The teacher says, “Listen to me”; “Watch me”; “My turn.” The teacher then shows students the skill that is being taught – what it looks like and sounds like.

Considerable evidence supports the role of teacher demonstration in the teaching of initial reading skills. Students of teachers who use explicit instruction to teach basic skills learn more than students who receive less explicit instruction such as discovery learning (e.g., Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Gunn, Smolkowski, Biglan, Black, & Blair, 2005; Klahr & Nigam, 2004; Kulik, Kulik, & Bangert-Drowns, 1990; Stallings, Robbins, Presbrey, & Scott, 1986). Modeling the appropriate response allows students to imitate the response, which is faster than trial-and-error learning and individual discovery (Meltzoff et al., 2009; Smith, 1979). Teacher demonstrations are also an effective instructional tool for teaching basic skills to young children who may be unable to learn new information or skills from a less-direct approach.

1.1.2. Student independent practice

Increased rates of independent practice are associated with the successful acquisition of new skills and a defining feature of proficiency in music, sports, and basic academic learning (Ericsson et al., 2007; Fields, 2005). For literacy acquisition, practice that targets word-level and reading fluency has been shown to improve comprehension (Vadasy, Sanders, & Peyton, 2005) independent of working memory (Swanson & O’Connor, 2009). Practice has also been shown to promote language acquisition (Meltzoff et al., 2009).

In the classroom, independent practice is conceptually similar to opportunities to respond (Council for Exceptional Children, 1987) and engagement in academic responding (Greenwood, Delquadri, & Hall, 1984), which refers to a combination of classroom behaviors including reading aloud, asking and answering questions, and participating in tasks. In early reading, where fluent and accurate word recognition is a key instructional goal, practice is the main vehicle by which young learners learn to decipher new words and sounds on their own.

The importance of student independent practice is based on evidence from several decades of research supporting the impact of opportunities to respond or academic engagement on students’ academic achievement (Greenwood, Horton, & Utley, 2002; Rosenshine, 1995). A brisk rate of independent opportunities for practice helps ensure that students’ attention is focused on the lesson (Carnine, Silbert, & Kame’enui, 1997; Gleason & Hall, 1991) and that they receive frequent opportunities to develop automaticity and fluency in basic skills, a prerequisite to skilled reading (Samuels, 1997; Share, 2008). Among students with moderate and severe disabilities, increasing rates of practice have been found to be an effective teaching practice for improving both academic and behavioral outcomes (Logan, Bakeman, & Keefe, 1997; Sutherland et al., 2003; Sutherland & Wehby, 2001). Similarly, researchers have found that the most effective general education teachers provide extensive practice to help their students to develop well-connected networks (Brophy & Good, 1986; Fields, 2005; Rosenshine, 1995).

1.1.3. Student errors

Student errors are defined as an incorrect response or no response during independent practice. Errors can happen for a number of reasons. The student may not understand what to do for the task, may have learned the skill incorrectly, or may not have learned the skill at all. Regardless of the reason, documenting student errors provides important information about the effectiveness of teachers’ instructional practices. We acknowledge that, in isolation, student errors do not represent an instructional interaction or approach per se, but we have chosen to document student errors because the rate of errors (a) gives an indication of the effectiveness of teachers’ instructional practices, (b) helps frame the interpretation of corrective feedback, and (c) provides an objective means for studying the relationship between error rates and reading gains.

1.1.4. Corrective feedback

Corrective feedback is defined as an instructional practice that directs students’ attention to their incorrect responses. To be corrective feedback, the teacher must provide some information about the task or skill (e.g., “remember, the letter makes the/mmm/sound”). Repeating a question or guiding the student to the correct answer is not corrective feedback (e.g., “What sound do these letters make?” or “Sound it out.”). Corrective feedback following a student error indicates that the teacher is monitoring students’ understanding during instruction and is responsive to their errors.

The significance of corrective feedback stems from research focused on comparisons of feedback techniques and the effects on word recognition in beginning readers (e.g., Barbetta, Heward, Bradley, & Miller, 1994; Meyer, 1982; Pany & McCoy, 1988). Findings from these studies suggest that the use of direct corrective feedback enhances word recognition accuracy and, in some cases, reading comprehension. An analysis of these studies conducted by McCoy and Pany (1986) found that corrective feedback was associated with more accurate word recognition and did not appear to interfere with comprehension during reading. Findings from both types of research also indicate that young children require more corrective feedback than those at a more advanced level of learning because they have not mastered the skills needed to automatically self-correct (Gardner, 1998).

1.2. Growing use of observational instruments

Observations of instruction inform researchers about real-time teaching practices and about the intervening classroom variables that affect changes in student outcomes. Findings from observations can be used to evaluate and improve the quality of instruction that students receive. Observational data also have the potential to advance the science of learning and build a closer link between research and classroom practice (Bransford, Brown, & Cocking, 2000).

Observation instruments are increasingly being utilized to document implementation of federally funded programs, such as Reading First (Baker et al., in press), and as an integral part of large-scale professional development systems. Widely used observation instruments measure multiple aspects of classroom instruction and provide useful information on the overall quality of early literacy teaching, including teacher affect, classroom settings, instructional design, content, and delivery. Global measures, such as the Classroom Assessment and Scoring Instrument (CLASS; Pianta, La Paro, & Hamre, 2008) and The Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 2004) assess the presence or absence of broad classroom features. For example, the CLASS documents classroom organization and emotional and instructional supports. Similarly, the ECERS-R measures physical arrangements, personal care routines, language use and reasoning-skill development, activities, program structure, staff–child interactions, and provisions for parent and staff in preschool and kindergarten classrooms. Content-specific observation instruments provide more detailed descriptions of instructional quality, specifically for the development of early literacy skills. The Teacher Behavior Rating Scale (TBRS; Landry, Crawford, Assel, Gunnewig, & Swank, 2004) documents teacher instructional practices related to written expression, print and letter knowledge, phonological awareness, book reading, oral language use, and mathematics concepts. The Early Language and Literacy Observations Tool (ELLCO; Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002) documents preliteracy activities, including storybook reading, circle-time conversations, and children's story writing. Connor et al. (2009) have attempted to provide further detail on early literacy teaching by documenting how dimensions of instruction, such as explicit versus implicit, teacher- versus child-managed, and changes in type and amount of instruction, interact with child outcomes on measures of vocabulary and decoding.

Although current global quality and content-specific observation instruments provide useful information on many facets of literacy instruction, there is still more to learn about how to effectively teach all children to read. Despite significant growth in the knowledge base on reading acquisition (Adams, 1990; Barnett, Robin, Hustedt, & Schulman, 2003; Cooper, Masi, & Vick, 2009; National Reading Panel, 2000; Snow, Burns, & Griffin, 1998; Vellutino et al., 1996), many children in the U.S. still leave the primary grades without the foundational literacy skills they need to succeed in school (Lee & Burkam, 2002).

We suggest that instructional delivery, characterized by student–teacher interactions, is an aspect of early literacy instruction that has received limited research attention but is an alterable instructional variable that has the potential to improve early reading achievement. Neither global quality nor content-specific instruments document the finer-grained interactions between teachers and students that may have a unique and direct effect on how well students learn to read. Nor can most current instruments provide teachers with feedback on ways to improve their delivery of early literacy instruction. To that end, we believe there is an important need to now focus on reliably measuring and more closely studying the instructional interactions that may develop basic literacy skills using an instrument expressly designed for that purpose.

1.3. Classroom Observations of Student–Teacher Interactions (COSTI)

The COSTI documents the rate of four student–teacher interactions that we believe are most important during beginning reading instruction and are not measured by other instruments: teacher demonstrations, independent student practice, student errors, and teacher corrective feedback. These interactions, alone and in combination, commonly occur during beginning reading instruction and have been recommended for basic skills instruction in many curricula and interventions. They have also been studied in single-case designs and with small samples of students with disabilities (e.g., Greenwood et al., 2002; Sutherland et al., 2003). Finally, the rate of these interactions can feasibly be altered by teachers to improve their teaching and thus their students' acquisition of basic early literacy skills.

Counting these observable, measurable interactions with clear decision rules permits observers to document teachers' instructional practices objectively, with fewer judgment calls that may affect the reliability and interpretability of results. Recording the duration of the observation allows for calculation of rates of interactions. Rates of interactions allow for more suitable comparisons when the duration of the observation sessions vary. Recording the frequency and rate of instructional interactions sequentially provides information that cannot be generated by other instruments because it documents the real-time occurrence of specific, observable, and malleable instructional practices in the order that they occur.

For the COSTI coding, we provided a more narrow definitions of student independent practice and corrective feedback, however, than used in some of the supporting literature. We define student independent practice as students' verbal practice of a task or skill such as sounding out a word or giving an answer to a question (e.g., "What sound does *m* make?") without any help from the teacher. We also distinguish independent practice from guided practice, in which the teacher asks students to practice but then provides help (e.g., leads the answer or practices with the students). During guided practice, the teacher may be unable to determine if students know the skill and can do the task independently because students may be able to perform the task only when they follow the teacher's lead. Observers use the COSTI instrument to code corrective feedback if a student gives an incorrect response or does not reply and the teacher immediately demonstrates the skill again, gives the correct answer, or gives part of the correct answer. These decision rules help reduce judgment calls and allow for more consistent and clearer interpretation of summary information.

1.4. Purpose of the COSTI

The COSTI serves two independent purposes: to provide teachers with feedback on their instruction in basic skills and to document student–teacher interactions in order to study effective instructional practices. The original version of the COSTI was developed and used by one of the authors (Gunn) to observe first-grade student teachers and to provide them with feedback on their interactions with students during reading instruction. These observations of interactions allow a coach to show teachers the patterns in their instruction, help them think about how those patterns may affect their student's learning, and suggest changes in their interactions during instruction. For example, a teacher may find that when he does not provide explicit demonstrations of new skills, his students make frequent errors, but when he provides consistent demonstrations, his students' understanding increases and rate of errors decreases. Similarly, a coach could use COSTI data to show a teacher that her pacing of independent student practice is slow and suggest that increasing the rate of practice may accelerate her students' learning, improve their task engagement, and decrease their disruptive behavior (Sutherland et al., 2003). Finally, a consistent pattern of errors that are not followed by corrective feedback may help explain to a teacher why some students "just aren't getting it." The ability to show objective data in such situations is more compelling to teachers than general impressions and helps lead to improvements in classroom instruction.

Subsequent versions of the COSTI have been field-tested in kindergarten and first-grade classrooms to refine the instrument and observer training and to make the tool one that can be feasibly and reliably used by teachers, coaches, and administrators. Data generated by the COSTI provide a practical, objective way to give teachers direct feedback on their instructional delivery. In this capacity, the COSTI serves to guide coaching and professional development directly aimed at helping teachers improve the quality of their instruction. See Exhibit 1 in Online Supplemental Appendix for an example tally sheet for coaching situations.

For the study of effective instructional practices, the instrument provides valuable, detailed information for researchers about an otherwise unmeasured aspect of classroom instruction in basic skills. For example, researchers seeking to test the efficacy or effectiveness of curricula may hypothesize that the rate of teacher demonstrations or student independent practice represents an important pathway to skill acquisition (e.g., Clarke et al., 2011; Gunn, Smolkowski, & Vadasy, 2011). In such investigations, the COSTI offers the only technically adequate measure currently available. The instrument can also be used to document student–teacher interactions across content areas in which basic skills are taught. Although this paper focuses on using the COSTI to study beginning reading instruction, the measure has also been used to examine the role of student–teacher interactions with early elementary reading programs in first grade (Baker et al., 2010) and kindergarten mathematics curricula (Doabler et al., 2010), thus extending its research applications. See Exhibit 2 in Online Supplemental Appendix for a generic, annotated example of a coding for research purposes.

1.5. *Research goals*

To demonstrate that the COSTI captures important instructional variables and serves its dual roles for professional development and research purposes, this paper aims to establish that (a) observers can reliably code the instructional interactions, (b) teachers are sufficiently stable in their behavior that a reasonably small number of observations can characterize their instructional interactions, and (c) the rates of student–teacher interactions predict student reading outcomes.

Our first goal was to demonstrate that, due to the simplicity and specific foci of the COSTI, trained observers could reliably document specific behaviors occurring between individual teachers and their students during beginning reading instruction. For this paper, we tested interobserver reliability, a basic quality critical to any observation tool, rather than agreement, which can be ambiguous and misleading (Rosenthal & Rosnow, 2008).

Classroom observations capture at least two additional sources of variation beyond interobserver reliability. Teachers' instructional practices vary from one occasion to another and teachers differ from each other. Observations of teacher behavior should describe the classroom context consistently over time. Our second goal was to test whether the teaching behaviors remain stable within teachers across the year. If not, more observation occasions would be required to characterize the behavior of interest adequately (Shoukri, Asyali, & Donner, 2004) and the instrument would be less useful. With the COSTI, we expected the rate of observed student–teacher interactions to remain relatively consistent across time within a classroom.

Our third goal was to demonstrate that the interactions measured by the COSTI would account for gains in student reading outcomes from the beginning to the end of kindergarten. If, for example, the observation instrument reliably captured independent student practice attempts, if teachers' instructional behaviors were consistent across time, and if practice had the value for student learning anticipated by its theoretical foundation, then we would expect to find that teachers who were observed to provided more practice would have students who made greater gains on reading measures across the school year. We also expected the observed interactions to have the largest relationships with measures of skills taught during kindergarten. For example, children learn much of their vocabulary at home, and most of the students in this study already knew at least 20 letter names by entry into kindergarten. This implies that kindergarten teachers would likely have a smaller relative impact on vocabulary and knowledge of letter names than on other early reading skills. We therefore expected the observed behaviors, and opportunities for independent practice in particular, to predict gains in reading decodable words, sight words, and phonological processing more than for vocabulary or knowledge of letter names.

2. Method

This paper reports on the reliability and validity of the COSTI collected within a school-randomized trial designed to test the *Read Well Kindergarten* (RWK) curriculum (Gunn et al., 2011). The trial entailed the use of RWK in approximately half of the schools, collection of student reading measures at the beginning and end of the school year, and observations of instructional behavior at three times during kindergarten. We designed the study to test the efficacy of the RWK curriculum and evaluated the COSTI within the same set of schools.

2.1. *Study design*

We recruited 24 elementary schools and all of their kindergarten teachers who agreed to participate. The schools were randomly assigned to teach the RWK curriculum or their usual literacy curriculum. All students for whom we had parent consent were tested in the fall and spring of kindergarten. Observations were collected on teachers in intact classrooms and student measures were nested within classrooms. Although the efficacy study represents a group-randomized trial (Murray, 1998) with schools assigned to condition, the present evaluation of the observation instrument treats classrooms or teachers as the primary units of analysis.

2.2. *Participants and procedures*

The study took place in 19 schools in rural and suburban communities in Oregon and five in rural New Mexico. Within each school, we followed two cohorts of kindergarten students through each classroom. The present paper does not distinguish between the two cohorts of students because we had no theory that predicted differences in the reliability or validity of the observed instructional interactions between them.

2.2.1. *Teacher and student sample*

This study included observations on 54 teachers, 26 in the RWK condition and 28 in the control condition. Only one teacher elected not to participate in the study. Within those classrooms, we assessed 1519 kindergarten students in the fall (T_1) and 1427 in the spring (T_2), with 50% in each condition. All students spoke English, although a few students spoke Spanish, Vietnamese, Mandarin, Arabic, or Tiwa (a Pueblo language) as their first language. Students in New Mexico were predominantly Hispanic, 60–80% by school, and 65–94% of the students received free or reduced lunch. Oregon students were predominantly White, 70–85% by school, and 20–60% of the students received free or reduced lunch. All kindergarten classrooms in New Mexico were full day. Four classrooms in Oregon were full day and the rest were half day for a full week or full day for 21/2 days per week. Table 1 provides additional information about the student, teacher, and school sample.

2.2.2. *Literacy instruction*

In RWK schools, teachers taught basic decoding, vocabulary, and comprehension skills through daily 20- to 25-min mastery-based lessons with three to four groups of two to eight students. Students moved through the units at a pace that was appropriate for them, and moved to the next unit when they met the passing criteria on the end-of-unit mastery test. All teachers received training from a certified *Read Well* trainer.

Teachers in comparison schools used commercially published kindergarten reading programs, most commonly *Harcourt Brace*, *McGraw Hill*, and *Houghton-Mifflin*, as well as *Zoo Phonics*, *Scott*

Table 1
Descriptive statistics for students, teachers, and schools by condition.

Measure	*Read Well Kindergarten*	Control schools
Students		
Age, *M* (*SD*)	5.7 (0.34)	6.2 (0.32)
Male	49.4%	50.9%
L1		
Spanish	3.3%	5.1%
Other	8.9%	6.5%
IEP		
Total	10.7%	11.5%
Speech and language	8.6%	8.6%
Autism	0.8%	0.7%
Specific learning disability	0.3%	0.9%
Developmental delay	0.6%	0.0%
Sample size	765	754
Teachers		
Years teaching, *M* (*SD*)	12.6 (10.12)	11.6 (10.10)
Years teaching K, *M* (*SD*)	4.9 (4.02)	5.5 (5.53)
Full day class	64%	57%
Sample size	26	28
Schools		
Free or reduced price lunch (%), *M* (*SD*)	52.2 (14.63)	54.6 (18.02)
Sample size	12	12

Notes. The sample sizes represent the maximum available. Age and sex were not available for 5 and 16 students, respectively. L1 refers to first language other than English. Language and individualized education plan (IEP) data were available for only 730 students in the *Read Well Kindergarten* (RWK) condition and 688 students in control classrooms. Teaching experience refers to the number of years at their first year of participation in this research project. One RWK teacher did not provide data on teaching experience.

Foresman, *Spalding*, *Animated Literacy*, *Celebrate Reading*, *Reading Milestones*, *Explode the Code,* and *Focusing on Language and Academic Instructional Renewal*. Some teachers supplemented their instruction with commercially produced or teacher-made literacy activities. Comparison group sizes ranged from one to 24 students (80% between 4 and 16).

2.2.3. Observation procedures

Trained project staff observed an entire literacy period on a regular school day in the fall, winter, and spring of each school year. Teachers provided basic literacy instruction to students grouped according to their instructional needs, so staff observed the instruction in one group at each time point, observing each group within each classroom at least once over the course of the study. Teachers participated across either 2 years ($n = 25$), contributing six observations, or 1 year ($n = 29$), with three observations. One teacher provided only five observations across 2 years and one teacher provided only two observations during 1 year. In 2005, we collected 66 observations on 22 teachers; in 2006, 115 observations on 39 teachers; and in 2007, 54 observations on 18 teachers. The 26 teachers in the RWK condition provided 113 observations and the 28 control teachers provided 122 observations. Observations took, on average, 23.5 min, and ranged from 4 to 93 min, although 90% of them fell between 10 and 44 min. The shorter times represent a few occasions where reading instruction was interrupted early. On average, observers coded instruction of seven students per observation occasion, and the number of students ranged from one to 24. On 24 occasions, a second observer collected data at the same time to test agreement and interobserver reliability.

2.2.4. Observer training

Observers were trained in four stages. First, observers were given an overview of the system, an explanation of the codes, and procedures for using the observation codebook as a reference. Second, the trainer and the observers practiced coding and debriefing observations as a group using video clips. Third, the observers practiced coding with the trainer in kindergarten classrooms that were not in the study. Fourth, reliability was established in project classrooms between the trainer and individual observers and maintained via periodic retesting. Observers that met and maintained an 80% or higher rate of agreement with the trainer were allowed to observe project classrooms.

2.3. Description of the COSTI form for coding instructional interactions

The COSTI observation form has a cover page where observers recorded general context information for each observation. For this study, we used the cover page to document grouping (i.e., whole class, small group), the number of students, the reading program or materials used, overall start and stop times, and masked identification information, such as ID numbers. The rest of the COSTI observation form was divided into 14 identical half-page sections. Because kindergarten reading instruction often targets multiple domains within one lesson, such as phonological awareness, alphabetic understanding, and comprehension, observers used a separate section on the form to code the instructional interactions for each activity within a lesson with a different instructional focus. An analysis of the COSTI by activity, however, was beyond the scope of the present manuscript. See Exhibit 3 in Online Supplemental Appendix for the observation form used for this project. We have also included the associated codebook as Exhibit 4.

2.3.1. Serial coding: how COSTI documents student–teacher interactions

Each half-page section of the observation form has a series of columns of bubbles to record the student–teacher instructional interactions, which are the main focus of this paper. Observers recorded, in sequence, each instance of (a) explicit teacher demonstrations, (b) student independent practice, (c) errors, and (d) teacher corrective feedback by filling in a bubble on the row allocated for that interaction. As noted in the introduction, this method of coding can be used for any reading activity in which students and teachers have observable, measurable, interactions. For example, in a demonstration the teacher would show students what the skill looks and sounds like: "Listen, the first sound in *ran* is/rrr/," or the teacher does what she wants the students to do "Watch me write the letter capital letter *J*." In student independent practice, the teacher asks students to provide some information that demonstrates their understanding of the activity: "Why did the children in this story get wet?" or "What sound does *r* make?"

This serial method of coding provided a total count of each interaction and the sequence in which the interactions occurred. We used these data from the COSTI to identify patterns of instructional interactions, such as whether a teacher demonstration was followed by independent student practice, whether that response was an error, and if so, whether that was followed by another teacher demonstration, a student practice, or a teacher correction. Because the duration of each observation session varied, a longer observation typically provided more interactions to code. Therefore, we also computed rates per minute for each instructional interaction to remove the influence of observation duration from the analyses. For example, in 20 min of instruction, an observer may code 10 teacher demonstrations, 100 independent student practice attempts, five errors, and two corrections. Dividing those counts by 20 min provides a rate per minute, such as .50 teacher demonstrations per minute or five student practice attempts per minute.

The data collected with the COSTI thus provided a quantitative summary of the instructional patterns that occur between students and their teachers during instruction in basic skills. It is important to recognize that the critical feature of the COSTI is the act of recording the four instructional interactions in a way that one can later calculate rates of interactions and tease out interesting patterns (e.g., that a correction directly followed an error). The particular style and layout of the data-collection form, however, is not critical to the interpretation.

2.4. Student measures

Trained project staff collected student reading performance measures during October and again in May, with some measures collected only in the spring (e.g., oral reading fluency).

2.4.1. Assessment training and procedures

Assessors included research assistants and retired or substitute teachers who were unaware of schools' intervention status. Assessors attended two 3-h sessions of standardized training prior to testing students. During the first day of testing in schools, the assessment coordinator first demonstrated how to test a student in order to observe and provide feedback on the fidelity and standardization of testing procedures.

2.4.2. Letter names and sounds

Students were shown a sheet with the uppercase letters of the alphabet in random order and asked to name the letters they knew. If students missed five letters consecutively, the examiner asked if they could name any other letters on the page. The score was the total number of letters named correctly. The same procedure was followed for asking students to identify letter sounds. The test was untimed and the maximum correct was 26 on each measure. These tests were given at pretest (names $M = 17.11$, $SD = 9.06$; sounds $M = 9.26$, $SD = 8.99$) and posttest (names $M = 24.28$, $SD = 3.96$; sounds $M = 20.93$, $SD = 6.11$). Although assessors were carefully trained and monitored, reliability estimates were unavailable. As a proxy, we conducted test–retest correlations across the school year as a lower bound on reliability. Letter name and letter sound scores correlated .55 and .46 from T_1 to T_2.

2.4.3. Sight words and decodable words

These two tests are curriculum-based measures that included sight words and decodable words taught in the phonics sequence of RWK. Sight words are words such as *was* or *said* that do not follow regular letter–sound relationships. Students were asked to read 10 commonly used sight words. Decodable words are words that follow regular letter–sound relationships such as *had* or *seem* and can be sounded out using letter–sound correspondences. Students were asked to read 12 words that are commonly found in beginning readers. Scores were reported as the number of words read correctly. Both tests were given at pretest (sight $M = 1.59$, $SD = 2.28$; decodable $M = 1.10$, $SD = 2.69$) and posttest (sight $M = 6.10$, $SD = 3.18$; decodable $M = 6.50$, $SD = 4.08$). Reliability estimates were unavailable, so we again conducted test–retest correlations across the school year with control students. Sight and decodable words correlated .46 and .43, respectively, from T_1 to T_2.

Project staff also administered two standardized measures of sight words and decodable words. The Woodcock Reading Mastery Test-Revised (Woodcock, 1998) is an individually administered battery of tests that measures multiple aspects of reading ability. The Word Identification (Word ID) subtest measures a student's ability to read sight words of increasing difficulty presented in isolation. The Word Attack subtest measures a student's ability to sound out and read a list of nonwords also presented in isolation. Because of the possible floor effects of these subtests when given in the fall of kindergarten, they were administered at posttest only. Internal consistency for the subtests ranges from .92 to .98, Word ID ($M = 11.86$, $SD = 14.21$) and Word Attack ($M = 6.26$, $SD = 6.80$). The analyses used raw scores.

2.4.4. Oral reading fluency

Oral reading fluency (ORF) is a valid, reliable indicator of overall reading competence (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Slocum, Street, & Gilberts, 1995; Stanovich, 2000). ORF was assessed only at posttest with a beginning first-grade decodable reading passage entitled "Mac Gets Well" (Makar, 1995). Eighty percent of the 165 words in the passage were decodable; the remaining were common sight words (e.g., *said, is, the, and, made*). Fluency rate was calculated as the number of words correctly read in 1 min ($M = 15.46$, $SD = 20.86$). Good and Kaminski (2002) show alternate form reliabilities between .89 and .94. Baker et al. (2008) reported test–retest reliabilities of .94 or higher and concurrent validity of .82 with the first-grade Stanford Achievement Test, Tenth Edition.

2.4.5. Phonological processing

The Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999) is a normed, comprehensive measure of phonological processing. The 20-item Phoneme Elision subtest measures the extent to which students can say a word and then say what remains after dropping out designated sounds (i.e., part of a compound word, part of an onset-rime, or an individual phoneme). The Elision subtest was given in the fall and spring to measure changes in students' phonological sensitivity. The test manual reported test–retest reliability of .88 for 5- to 7-year-olds. The CTOPP was given at pretest ($M = 2.43$, $SD = 2.69$) and posttest ($M = 5.02$, $SD = 3.85$). The analyses used raw scores.

2.4.6. Vocabulary

The Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997) is an individually administered, norm-referenced, wide-range test of receptive language or listening vocabulary. The test estimates English vocabulary acquisition and verbal ability. It requires students to select the one picture out of four that best illustrates the meaning of a stimulus word presented by the examiner. Students received Form IIIA at T_1 and Form IIIB at T_2. The PPVT has a mean of 100, standard deviation of 15, and test–retest reliability, as reported in the manual, of .92 for 2- to 6-year-olds and .93 for 6- to 10-year-olds. The PPVT was given at pretest ($M = 76.71$, $SD = 17.72$) and posttest ($M = 87.77$, $SD = 17.72$). The analyses used raw scores.

2.5. Statistical estimation methods

The statistical methods have been incorporated into the results to improve readability. We fit the multilevel models, described below, with PROC MIXED within SAS version 9.1 (SAS Institute, 2005). The multilevel models used restricted maximum likelihood, generally recommended for multilevel models (Singer & Willett, 2003; Snijders & Bosker, 1999), and assumed independent and normally distributed errors. We addressed the more important independence assumption (van Belle, 2002) using multilevel statistical models, and like regression and ANOVA, multilevel models have been found quite robust to violations of normality in a variety of scenarios (Donner & Klar, 1996; Fitzmaurice, Laird, & Ware, 2004; Hannan & Murray, 1996; Maas & Hox, 2004; Murray et al., 2006).

The analysis included only complete cases, but we believe that missingness did not bias the results of this study. Observation data were missing only twice out of the 237 opportunities (0.8%). With 1519 students at T_1 and 1427 at T_2, the prediction models included 1404 (92.4%) of the T_1 cases. While a large proportion of missing data can have a sizable impact on results (e.g., Smolkowski, Danaher, Seeley, Kosty, & Severson, 2010), an analysis with less

Table 2
Descriptive statistics for observation measures across all observation occurrences.

Measure	Frequencies				Rates per minute				*n*
	M (*SD*)	Percentile			*M* (*SD*)	Percentile			
		25th	50th	75th		25th	50th	75th	
Teacher demonstrations	17.9 (18.6)	4.0	13.0	27.0	0.8 (0.8)	0.2	0.6	1.2	235
Practice opportunities	133.0 (96.5)	54.0	108.0	204.0	5.9 (4.2)	2.4	5.3	8.5	235
Student errors	7.2 (7.5)	2.0	5.0	11.0	0.3 (0.4)	0.1	0.2	0.5	235
Teacher corrections	3.3 (4.3)	1.0	2.0	4.0	0.2 (0.2)	0.0	0.1	0.2	235
Practice opportunities followed by an error	5.4 (5.9)	1.0	4.0	8.0	0.3 (0.3)	0.1	0.2	0.4	235
Errors followed by a teacher correction	3.1 (3.9)	1.0	2.0	4.0	0.1 (0.2)	0.0	0.1	0.2	235

Notes. The sample size, *n*, represents the number of observation occurrences. The average observation duration was 23.5 min, and the average number of students present during the observation was 6.9.

than 8% missing should be relatively unbiased (Schafer & Graham, 2002).

3. Results

We provide descriptive, reliability, and variability information for the sequential measures using count, rate, and proportion metrics. We then provide results of the validity analyses, where we focus on the more theoretically interesting rate-per-minute and proportion measures. See Online Supplemental Appendix for additional descriptive information, results from tests of the sensitivity of the measures to differences between curricula, and detailed results from the predictive validity models.

3.1. Descriptive results

Table 2 provides the means, standard deviations, quartiles, and sample sizes of the directly observed measures and two conditional variables for the full sample. The table shows the descriptive information for both frequencies and rates. We observed a minimum of zero occurrences for each instructional measure at least once, except for practice opportunities, which had a minimum of .40 opportunities per minute, and for at least one teacher, we observed no demonstrations during the year.

We computed partial correlations among observation measures that controlled for treatment condition because the data were collected within an active treatment study. Table 3 lists the partial correlations among a subset of observation measures when controlling for treatment condition. The table presents the results for all 235 observations and for means across the three observations within the 79 classrooms. The high correlations between practice, errors, corrections, and errors followed by a correction were not surprising as they were structurally related within the observation system; the rate of errors followed by a correction was excluded from subsequent analyses due to this dependence. Overall, the partial correlations in Table 3 were only slightly lower than correlations that did not control for condition.

3.2. Interobserver reliabilities

Interobserver reliability refers to the degree to which individual observers, coding the same set of behaviors in the same classroom, provide the same information. Our model assumes that codes differing between pairs of observers watching the same teacher represent an observer error while agreement represents the “true” student–teacher interactions. To test reliability, we computed intraclass correlation coefficients (ICC; McGraw & Wong, 1996; Shrout & Fleiss, 1979) or generalizability coefficients (Mitchell, 1979) from multilevel models with observers nested within observation occasions. The models provided two variance estimates: the error variance corresponds to differences between observers within observation occasions, and the classroom variance constitutes variation between teachers. The ICC gives the proportion of total variance at the classroom level. Highly reliable observers, both watching and coding the same set of events, will provide very similar information, leading to little observer variation relative to classroom-level variation and a large ICC. Landis and Koch (1977) provided guidelines for the interpretation of ICCs: slight reliability, 0.00–0.20; fair, 0.21–0.40; moderate, 0.41–0.60; substantial, 0.61–0.80; and nearly perfect, 0.81–1.00 (see also Mitchell, 1979).

In our sample, two observers simultaneously collected data on 24 of the occasions (10%) in order to test reliability. Given the high rates of agreement in preliminary studies, we expected that this sample size would be adequate for the analyses (Shoukri et al., 2004; Walter, Eliasziw, & Donner, 1998). Table 4 presents the ICCs as a measure of the reliability of observers for the four interactions that were directly observed and coded. The interobserver reliabilities for the rate and frequency of student practice opportunities was particularly high, and all ICCs, which ranged from .61 to .99, represent substantial to nearly perfect reliability (Landis & Koch, 1977).

The reliability values were lower for the rate-per-minute measures than frequencies for each measure. This resulted from a difference in the start and stop times between observers in one instance. One observer reported an observation time from 09:35 to 09:56 while the other reported 09:37 to 09:44. From an examination of the times for individual activities, it is likely that the 9:56 stop time should have been coded as 9:46. Correcting that time value improved reliabilities of the rate measures (e.g., from .855 to .978 for student practice opportunities) but had a negligible impact on other results. We presented the results with the original time, however, because we could not verify that a mistake occurred.

3.3. Classroom and teacher stability

We next tested whether interactions remained stable within teachers across the year. This is important because the lack of stability would require more observation occasions across the school year to adequately characterize the behaviors of interest (Shoukri et al., 2004). We again used an ICC, but calculated it differently than for interobserver reliability.

To demonstrate stability of observed behaviors across the school year, we fit a model with the three observations per classroom nested within each of the 79 classrooms during each school year. The within-classroom, Level 1, or error variance provides an estimate of the day-to-day variability in instructional activities plus any unreliability in the measure. The between-classroom or Level 2 variance estimates the difference between different teachers' instructional styles and between different years for teachers who participated twice. We used the ICC here as an estimate of the proportion of between-teacher variation and to help determine the need for more or fewer observations (Donner & Eliasziw, 1987;

Table 3
Partial correlations controlling for treatment condition among observation measures for all observations ($n = 235$) in the upper right and all classrooms ($n = 79$) in the lower left.

Measure		1	2	3	4	5	6
Demonstrations per minute	1	–	.23*	.14*	.18*	.17*	.01
Student practices per minute	2	.26*	–	.38*	.26*	.24*	−.15*
Student errors per minute	3	.45*	.44*	–	.75*	.75*	.59*
Teacher corrections per minute	4	.39*	.29*	.79*	–	.98*	.44*
Errors followed by a correction per minute	5	.36*	.27*	.78*	.99*	–	.45*
Proportion of practice followed by an error	6	.50*	−.18	.56*	.48*	.47*	–

* $p < .05$.

Shoukri et al., 2004). High ICCs indicate stable behavior and imply that fewer observations will adequately capture the behaviors of interest. Lower ICCs imply the need for more observations to obtain a reasonable estimate of teacher behavior across the school year. The results of Shoukri and colleagues suggest that ICCs greater than .50 need no more than three observations per teacher over the period of interest, such as the school year. ICCs in the .20 to .50 range would require between three and six replicates per teacher. Because ICCs can be interpreted as the average correlation between given pairs of observations for the same teacher across a school year (Shrout & Fleiss, 1979), the stability ICCs reported here are analogous to test–retest reliability estimates.

Table 4 also presents the ICCs that describe teacher stability over time. The ICCs ranged from below .10 for teacher demonstrations to above .50 for practice opportunities. These stability estimates imply that the rate of demonstrations is not very stable. In contrast, with an ICC of .59, the rate of independent practice opportunities is stable enough over the school year that three observations per year should sufficiently capture the rate of practice opportunity interactions. Within these models, we also tested for a generally increasing or decreasing trend across the year by adding a linear time component (Fitzmaurice et al., 2004). This analysis failed to find evidence of a trend.

3.4. *Reliability of the observed aggregate means*

In studies that collect observation data repeatedly, investigators often aggregate measures of observed behavior for each classroom for analyses or reporting (e.g., Gunn et al., 2011). The *reliability of the observed means* (Snijders & Bosker, 1999) offers the best estimates of the reliability of such aggregated data. The reliability of the observed means also corresponds to the empirical Bayes estimate provided by hierarchical linear models (Raudenbush & Bryk, 2002). We estimated the reliability of the mean of the three observations per classroom. The calculation of the reliability is similar to that for the ICC (Snijders & Bosker, 1999). If we represent the classroom-level variance with τ^2, the within-classroom or observation-level variance with σ^2, and the number of observations in classroom j as n_j, then while the ICC equals $\tau^2/(\tau^2 + \sigma^2)$, the reliability of the observed mean equals $\tau^2/(\tau^2 + \sigma^2/n_j)$. The reliability of the observed mean with one observation per classroom equals the ICC, and the reliability of the mean will improve as the number of observations per teacher increases. The reliability of the observed mean provides upper limits on the potential validity value of each measure when averaged across the year.

Table 4 also presents estimates of the reliability of the observed means. Except for teacher demonstrations, the reliabilities of the serially coded measures for classrooms ranged from .45 to .88, fair to excellent. The mean of the number and rate of teacher demonstrations were not particularly reliable measures of the "true" classroom-level interactions. Additional observations during the school year would have improved the reliability of these means.

3.5. *Prediction of student outcomes*

To demonstrate validity, we tested whether the interactions measured by the COSTI would account for gains in student reading outcomes during kindergarten. We focused on the rate-per-minute measures and the proportion of independent practice opportunities followed by an error. The analysis involved a series of multilevel models that nested covariate-adjusted student scores within classrooms and then predicted the classroom-level student scores with the mean of the three observations across the school year (Singer & Willett, 2003). The statistical model can be represented by equations for the student and classroom levels, respectively:

$$\begin{aligned} Y_{ij} &= \pi_{0j} + \pi_{1j}X_{ij} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2) \\ \pi_{0j} &= \beta_{00} + \beta_{01}O_j + r_{0j} \qquad r_{0j} \sim N(0, \tau^2) \end{aligned}$$

The student-level equation predicts each student score, Y_{ij}, with a classroom intercept, π_{0j}, an individual-level covariate effect,

Table 4
Interobserver reliabilities, teacher stability over time, and reliabilities of observed means for the observed student–teacher interactions.

Measure	Unit	Interobserver reliability	Teacher stability over time	Reliabilities of the observed means
Teacher demonstrations	Frequency	.631	.086	.221
	Rate per minute	.609	.095	.239
Student practice opportunities	Frequency	.993	.710	.880
	Rate per minute	.855	.588	.811
Student errors	Frequency	.776	.405	.672
	Rate per minute	.718	.324	.590
Corrections	Frequency	.679	.284	.544
	Rate per minute	.683	.214	.449
Proportion of practice followed by an error	Proportion		.093	.236

Notes. The interobserver reliabilities, an intraclass correlation (ICC), were computed with 24 observation occasions that were each coded by two observers where each of the two sets of observation data were nested within each observation occurrence. The ICC represents the teacher-level or "true" variance. Estimates of teacher stability over time were also described by an ICC and used the entire sample of 79 classrooms. The three observations within each classroom collected across the school year were nested within the classroom, and the ICC describes the proportion of variation that was consistent across the three observations within a classroom over the school year. Estimates of the reliability of the observed means demonstrate the reliability of an aggregated observation variable based on all three observations collected in each classroom during the school year.

$\pi_{1j}X_{ij}$, and a student-level error, e_{ij}, with variance σ^2. Covariate-adjusted outcomes account for individual differences at pretest, before any instruction in kindergarten took place, allowing the prediction of student improvement on literacy measures over the school year. For covariates, we used the same measure collected at pretest when available and letter names and letter sounds otherwise.

The classroom-level equation predicts the adjusted student mean, π_{0j}, with an intercept, β_{00}, the effect of an observation measure, $\beta_{01}O_j$, and the classroom-level error, r_{0j}. The classroom intercept describes the classroom mean of covariate-adjusted student performance expected when the observation measure equals zero. A statistically significant β_{01} term indicates that the observation measure accounts for variation in student outcomes beyond chance. Tests for the β_{00} and β_{01} have 77 degrees of freedom (*df*) as we have 79 classrooms. The pretest covariates at the student level do not influence the *df* associated with tests of the β_{00} and β_{01} terms.

To ease interpretation of effects, we computed two effect sizes for each predictor. The first was a partial correlation coefficient, r_{partial}, converted from the *t*-value and *df* associated with the test of O_j (Rosenthal & Rubin, 2003): $r_{\text{partial}} = t/\sqrt{(t^2 + df)}$. In models with one outcome and one predictor, r_{partial} provides an effect-size estimate equal to the usual Pearson correlation but at the classroom level, and in regression models with multiple predictors, r_{partial} gives the magnitude of the relationship between the outcome and predictor controlling for all other variables in the regression equation. The Pseudo-R^2 statistic (Singer & Willett, 2003) estimates the proportion of classroom-level variance reduced by the addition of each observation measure as a predictor. For an unconditional model *u* with no $\beta_{01}O_j$ term and a conditional model *c* with the $\beta_{01}O_j$ term, the classroom-level Pseudo-R^2 can be defined as follows: Pseudo-$R^2 = (\tau_u^2 - \tau_c^2)/\tau_u^2$. The interpretation of effect sizes from multilevel models, however, becomes complicated due to the multiple sources of variation (Singer & Willett, 2003; Snijders & Bosker, 1999). The reported r_{partial} statistic represents a classroom-level effect size and would not necessarily correspond to effects for individual students. Kraemer (2005) has also raised concerns about r_{partial}. Therefore, we report both r_{partial} and Pseudo-R^2 but recommend care when interpreting them.

Table 5 presents the validity of the COSTI measures with estimate of the Pseudo-R^2, the partial correlation coefficient, and the *p*-value associated with the test of statistical significance. The *p*-values estimate the probability that the relationship was at least as large as that expected under the null hypothesis (Hogg & Craig, 1978). In Table 5, some observed measures increased the between-classroom variation, signified by a negative Pseudo-R^2. In all such cases, the Pseudo-R^2 values were small and the predictors were not statistically significant. Finally, we tested the relationships between observed behaviors and outcomes while also controlling for treatment condition, observed class size, and duration of observation. The details differed, but the additional predictors did not appreciably change the pattern of results.

Overall, the rate of practice accounted for substantial variance in all student literacy measures except for the PPVT, with statistically significant results and Pseudo-R^2 values that ranged from .14 to .39. The other student–teacher interactions, however, did not predict student literacy outcomes as strongly. To illustrate the interpretation of the results, we provide a detailed example for the decodable words measure. The analyses included the pretest measure as a covariate, so the outcome can be thought of as the gains in decodable words across kindergarten. Table 5 shows that improvements in decodable word reading were statistically significantly predicted by the rates of student independent practice ($t = 6.35$, $p < .0001$; *t*-values not tabled), student errors ($t = 2.50$, $p = .0146$), and teacher corrections ($t = 2.20$, $p = .0311$), as well as the proportion of practice opportunities followed by an error ($t = -3.60$, $p = .0006$). Teacher demonstrations did not meaningfully predict gains in decodable word scores. The rate of opportunities for student practice reduced the between-classroom variance in the mean gains in decodable words read correctly by 39% (Pseudo-$R^2 = .39$). Similarly, the correlation between the rate of student independent practice and decodable words, controlling for the pretest covariate, was .59. The proportion of practice opportunities followed by an error, in contrast, had an inverse relationship with decodable words, indicated by the negative partial correlation. A lower proportion of independent practice followed by errors predicted gains on decodable words read correctly.

4. Discussion

The results presented above demonstrated that the student–teacher instructional interactions captured by the COSTI can be reliably coded, that independent opportunities for practice and some other interactions remain stable for teachers across the school year, and that the classroom-level means are reliable. The observed measures, particularly practice opportunities, also predicted a number of literacy outcomes. In this section, we summarize the support for the COSTI, discuss limitations, and present implications for research and practice.

4.1. *Results summary*

4.1.1. *Interobserver reliability*

We established reliability of the COSTI in two ways. First, observers were required to maintain an 80% rate of agreement or higher with the observation trainer before they were allowed to observe teachers in classrooms, and these agreement tests were periodically repeated throughout the project. The percent agreement allowed for quick tests of observer congruity during training and ongoing project activities, but it fails to demonstrate true reliability due to its dependence on the base rate of the behavior coded (Mitchell, 1979). Second, the analyses for this paper provided rigorous tests of reliability with estimates of ICCs (McGraw & Wong, 1996; Shrout & Fleiss, 1979), a chance-corrected measure of agreement (Fleiss & Cohen, 1973). Observers attained strong reliability estimates for independent practice opportunities, the measure of most interest. Observers also reliably collected data on teacher demonstrations, student errors, and teacher corrections, although improvements could be made for teacher demonstrations and corrections. These interactions overlap in some cases because teacher demonstrations may be used to correct a student error as well as for modeling new skills. Future training might focus additional attention on clearly discriminating between occasions when a demonstration has been used for new instruction and when it serves as a correction.

4.1.2. *Classroom and teacher stability*

We hypothesized that teaching behaviors such as teacher demonstrations and opportunities for students to practice new skills independently would remain stable within teachers. The results showed that one could generalize from 1 day to the next for some of the instructional interactions. Our highest ICCs were in the range of .40 to .71 for practice opportunities and student errors, which were sufficiently high to recommend two or three observations per year (Shoukri et al., 2004). Student independent practice was stable across the school year; teachers who tended to give more practice opportunities tended to do so consistently. Teacher demonstrations appear to vary considerably from one occasion to the next, but the low ICC may also be partially due to the lower reliability of the measure. Lower ICC values imply that the data depend somewhat on the day they were collected and would likely require

Table 5
Validity estimates – effect sizes and statistical significance levels – for the observation measures as classroom-level predictors of pretest-adjusted student literacy outcomes.

Measure	Student outcomes								
	Letter names	Letter sounds	CTOPP	Sight words	Decodable words	Word ID	Word attack	Oral reading fluency	PPVT
Demonstrations per minute									
Pseudo-R^2	.01	.01	.01	.02	.01	.06	.02	.06	−.04
Partial *r*	.20	.16	.17	.17	.15	.25	.18	.24	.08
p-Value	.0785	.1621	.1304	.1296	.1763	.0259	.1038	.0332	.4735
Student practice per minute									
Pseudo-R^2	.22	.21	.25	.36	.39	.14	.14	.17	−.04
Partial *r*	.32	.43	.36	.55	.59	.33	.35	.34	−.01
p-Value	.0043	<.0001	.0013	<.0001	<.0001	.0026	.0016	.0022	.9467
Student errors per minute									
Pseudo-R^2	−.00	.05	.13	.15	.07	.11	.07	.18	.02
Partial *r*	.12	.24	.26	.38	.27	.30	.26	.35	.16
p-Value	.3064	.0343	.0187	.0007	.0146	.0064	.0210	.0014	.1614
Teacher corrections per minute									
Pseudo-R^2	−.02	.02	.08	.11	.06	.03	.01	.07	−.02
Partial *r*	.06	.16	.21	.31	.24	.18	.14	.22	.09
p-Value	.6137	.1666	.0617	.0050	.0311	.1165	.2129	.0495	.4344
Proportion of practice followed by an error									
Pseudo-R^2	.03	.04	−.03	.05	.16	−.01	−.02	−.02	.05
Partial *r*	−.13	−.19	−.04	−.23	−.38	.10	−.03	.11	.22
p-Value	.2686	.0987	.7190	.0455	.0006	.3634	.7820	.3344	.0568

Notes. Significance levels, Pseudo-R^2 values, and partial *r* values produced from a multilevel model that nested covariate-adjusted student scores with classrooms and predicted the classroom-level student outcomes with classroom means of observed behaviors. Student outcomes were adjusted for pretest measures of letter names and letter sounds except for CTOPP, sight words, and decodable words. Tests of fixed effects used 77 degrees of freedom. The Pseudo-R^2 represents the decrease in between-classroom variance from the unconditional to each conditional model that includes an observation measure as a predictor. The partial *r* characterizes the magnitude of the relationship between the predictor and outcome variables controlling for all other variables in the model.

additional observations during the school year or procedures to improve reliability.

One can interpret these ICCs as the average correlation between randomly chosen pairs of observations for the same teacher (Shrout & Fleiss, 1979), so the correlation between the frequencies of practice opportunities observed on any two occasions would be about .71 within classrooms. This is notable given that the students frequently differed across the instructional groups observed for each teacher, and observations took place several months apart.

4.1.3. Reliability of the observed means

The reliability of observed means describes how well each observed variable, aggregated across the school year, represents the interactions of interest. It "measures the ratio of the *true score* or parameter variance, relative to the *observed score* or total variance of the sample mean" (Raudenbush & Bryk, 2002, p. 46, emphasis in original). In the present context, this is the reliability of the observation measures when averaged at the classroom level (Snijders & Bosker, 1999), which were the variables used in the predictive validity models herein, the tests of the effectiveness of the RWK curriculum (Gunn et al., 2011), and the variables likely to be used in future studies that predict student outcomes. The reliabilities of observed classroom means ranged from .45 to .88 for independent practice, student errors, and teacher corrections, and they were somewhat lower for teacher demonstrations and proportion of practice followed by an error. In sum, the reliabilities of the observed means demonstrated that the observation system provided an excellent measure of the classroom context for practice opportunities and an acceptable estimate for the measures of student errors and teacher corrections.

4.1.4. Prediction of student outcomes

Critical student–teacher interactions ideally reflect important antecedents to students' acquisition of new skills. We hypothesized that the COSTI, and particularly student independent practice, would account for gains in student reading outcomes from the beginning to the end of kindergarten. Overall, the rate of independent practice was the most consistent predictor of student outcomes, indexed by either effect sizes or statistical significance. The rate of practice accounted for over 35% (Pseudo-R^2) of the between-classroom variability in gains on the CBM measures of sight words and decodable words, 25% of the classroom variance of phonological processing, over 20% for letter names and sounds, and 15% of the variance in oral reading fluency. As expected, we found the largest relationships between the rate of practice and the measures of those skills taught during the kindergarten school year, such as reading decodable words, sight words, and phonological processing.

While the rate of student errors and teacher corrections accounted for some variability in student outcomes, these relationships are likely a function of their dependence on the rate of independent student practice. A lower proportion of practice with errors was associated with better outcomes for sight words and decodable words. We also found that the proportion of practice with errors was negatively associated with independent student practice. The rate of teacher demonstrations was not related to student outcomes. While demonstrations may be an important factor, their value as a predictor rests partially on their reliability and stability, which were both low (Salvia & Yssledyke, 1998). Teacher demonstrations may also require a more sophisticated assessment that addresses more than just the number or rate, such as the quality or sequence of the demonstrations (Engelmann & Carnine, 1991).

4.2. Study limitations

The findings reported in this paper include a number of limitations. First, the analyses stretched the limits of the sample – we conducted numerous analyses with 79 classrooms. This sample provides just enough data to avoid overfitting and spurious results (Babyak, 2004). The paper also presented a large number of statistical tests, and we did not adjust our criterion for statistical significance for capitalization on chance. Although we believe the results presented here paint a consistent picture and imply the importance of independent student practice for effective instruction, replication is paramount and with different students, teachers, settings, and content areas. Such replication has begun, but until data from the replication studies become available, we caution against overgeneralizing from the present results. When interpreting the results, we recommend a focus upon the strong and consistent outcomes for independent student practice. We did not evaluate the effects of student moderators, such as whether the predictive validity of the rate of practice opportunities might depend on students' initial performance or background characteristics (e.g., socioeconomic status). Investigation of these and other relationships with literacy outcomes would be informative. Finally, while we believe the COSTI can guide coaching and professional development aimed at improving the quality of instruction, use for this purpose would be better supported by tests of reliability and validity conducted with practitioners in the schools, such as school psychologists or literacy coaches, rather than trained research assistants. We believe, however, that due to the simplicity of the coding system, practitioners should be able to reliably use the COSTI to provide teachers with beneficial feedback on their instruction.

4.3. Implications and future directions

Overall, the results support the reliability, generalizability, and predictive validity of the COSTI. Additional results support the contribution of independent student practice to the development of beginning reading skills within evidence-based curricula. Gunn et al. (2011) found that the combination of the RWK reading program and high rates of independent student practice led to the greatest gains in literacy skills for kindergarten students. Although the idea that practice makes perfect is hardly new (Ericsson et al., 2007; Newell & Rosenbloom, 1981), it has received limited attention in the research on the acquisition of basic academic skills. For example, word recognition proficiency is a prerequisite for all literacy learning (National Reading Panel, 2000; Share, 2008). During grades K-2 when children learn the basic skills that support accurate, fluent word recognition, it appears to be essential that students receive frequent opportunities to practice those skills. Conversely, when students move beyond initial skill development to higher levels of knowledge and skill application, such as text comprehension, the appropriate instructional approaches and documentation of those approaches may take a different form. More detailed knowledge about how best to teach those skills and how to measure the teachers' instruction of those skills would be valuable to the field.

We suggest that COSTI provides information about instructional interactions not captured by other measures and may prove useful for intervention studies that attempt to increase student–teacher interactions (Pianta & Hamre, 2009; Raudenbush, 2008) or improve the quality of program implementation. Complemented with additional data, such as observations of the general classroom environment (Pianta & Hamre, 2009) or teacher logs (Rowan & Correnti, 2009), the COSTI may help provide a more comprehensive picture of classroom instruction. As the field moves toward an evidence-based model of education, such information could also inform the skills taught in teacher education programs and refine the emphasis in coaching and mentoring programs currently in place in schools across the country.

Acknowledgements

This manuscript was funded by the Institute of Education Sciences Field Initiated Studies Grant R305F050080. We would also like to thank Kathy Jungjohann and Doug Carnine for their mentorship and guidance on observing and coaching teachers that led to the development of this instrument. We would like to thank Susan Long for her expertise and assistance in preparing this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecresq.2011.09.004.

References

Adams, M. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Archer, A. L. & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching*. New York, NY: Guilford.

Aylward, E., Richards, T., Berninger, V., Nagy, W., Field, K., Grimme, A., et al. (2003). Instructional treatment associated with changes in brain activation in children with dyslexia. *Neurology*, *61*, 212–219. doi:10.1212/01.WNL.0000068363.05974.64

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*, 411–421. doi:10.1097/01.psy.0000127692.23278.a9

Baker, D. L., Linan-Thompson, S., Kosty, D. B., Smolkowski, K., Mielke, A. R. & Miciak, J. (2010, July). Reading intervention with Spanish-speaking students: Maximizing instructional effectiveness in Spanish in first grade. In *Poster presented at the 17th Annual Meeting of the Society for the Scientific Studies of Reading* Berlin, Germany.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J. & Thomas Beck, C. (2008). Reading fluency as a predictor of reading proficiency in low-performing high poverty schools. *School Psychology Review*, *37*, 18–37.

Baker, S. K., Smolkowski, K., Smith, J. M., Fien, H., Kame'enui, E. J., & Thomas Beck, C. The impact of Oregon Reading First on student reading outcomes. *Elementary School Journal*, in press.

Barbetta, P. M., Heward, W. L., Bradley, D. M. & Miller, A. D. (1994). Effects of immediate and delayed error correction on the acquisition and maintenance of sight words by students with developmental disabilities. *Journal of Applied Behavior Analysis*, *27*, 177–178. doi:10.1901/jaba.1994.27-177

Barnett, S. W., Robin, K. B., Hustedt, J. T. & Schulman, K. L. (2003). *The state of preschool: 2003 state preschool yearbook*. Camden, NJ: The National Institute for Early Education Research at Rutgers University.

Bransford, J., Brown, A. & Cocking, R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

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

Carnine, D., Silbert, J. & Kame'enui, E. (1997). *Direct instruction reading* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Carver, S. & Klahr, D. (2001). *Cognition and instruction: Twenty-five years of progress*. Hillsdale, NJ: Erlbaum.

Clarke, B., Doabler, C., Baker, S. K., Fien, H., Smolkowski, K. & Chard, D. (2011, February). Early Learning in Mathematics (ELM): Developing observation instruments to explore mediators of student achievement. In *Paper presented at the 2011 Pacific Coast Research Conference* San Diego, CA.

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

Cooper, J. L., Masi, R. & Vick, J. (2009). *Social–emotional development in early childhood: What every policymaker should know*. New York, NY: National Center for Children in Poverty, Columbia University Mailman School of Public Health.

Council for Exceptional Children. (1987). *Academy for effective instruction: Working with mildly handicapped students*. Reston, VA: Author.

Doabler, C., Fien, H., Smolkowski, K., Baker, S., Clarke, B., Kosty, D. & Strand Cary, M. (2010, March). Measuring instructional interactions in kindergarten math. In *Poster presented at the Third Annual Society for Research on Educational Effectiveness* Washington, DC.

Donner, A. & Eliasziw, M. (1987). Sample size requirements for reliability studies. *Statistics in Medicine*, *6*, 441–448. doi:10.1002/sim.4780060404

Donner, A. & Klar, N. (1996). Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, *49*, 435–439. doi:10.1016/0895-4356(95)00511-0

Dunn, L. M. & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Ebbinghaus, H., Ruger, H. A. & Bussenius, C. E. (1913). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College Press.

Engelmann, S. & Carnine, D. (1991). *Theory of instruction: Principles and applications* (Rev. ed.). Eugene, OR: ADI Press.

Ericsson, K., Roring, R. & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High Ability Studies*, *18*(1), 3–56. doi:10.1080/13598130701350593

Fields, R. D. (2005). Myelination: An overlooked mechanism of synaptic plasticity? *The Neuroscientist*, *11*, 528–531. doi:10.1177/1073858405282304

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619. doi:10.1177/001316447303300309

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C. & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, *90*, 37–55. doi:10.1037//0022-0663.90.1.37

Fuchs, L. S., Fuchs, D., Hosp, M. K. & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*, 239–256. doi:10.1207/S1532799XSSR0503_3

Gardner, H. A. (1998). *The role of error correction in working with emergent readers* (Report No. CSO13298). Orangeburg, NY: School of Education, Dominican College (ERIC Document Reproduction Service No. ED430207).

Gleason, M. M. & Hall, T. E. (1991). Focusing on instructional design to implement a performance-based teacher training program: The University of Oregon Model. *Education and Treatment of Children*, *14*, 316–332.

Good, R. H., & Kaminski, R. A. (2002). *DIBELS oral reading fluency passages for first through third grades* (Technical Report No. 10). Eugene, OR: University of Oregon.

Goswami, U. (2004). Neuroscience, education, and special education. *British Journal of Special Education*, *31*, 175–181. doi:10.1111/j.0952-3383.2004.00352.x

Greenwood, C. R., Delquadri, J. & Hall, R. V. (1984). Opportunity to respond and student academic performance. In W. Heward, T. Heron, D. Hill, & J. Trap-Porter (Eds.), *Behavior analysis in education* (pp. 58–88). Columbus, OH: Merrill.

Greenwood, C. R., Horton, B. T. & Utley, C. A. (2002). Academic engagement: Current perspectives on research and practice. *School Psychology Review*, *31*, 328–349.

Gunn, B., Smolkowski, K., Biglan, A., Black, C. & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education*, *39*, 66–85. doi:10.1177/00224669050390020301

Gunn, B., Smolkowski, K. & Vadasy, P. (2011). Evaluating the effectiveness of Read Well Kindergarten. *Journal of Research on Educational Effectiveness*, *4*(1), 53–86. doi:10.1080/19345747.2010.488716

Hannan, P. J. & Murray, D. M. (1996). Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed model and the logistic mixed model analyses in simulated community trials with a dichotomous outcome variable at the individual level. *Evaluation Review*, *20*, 338–352. doi:10.1177/0193841X9602000306

Harms, T., Clifford, R. M. & Cryer, D. (2004). *Early Childhood Environment Rating Scale* (Rev. ed.). New York, NY: Teachers College Press.

Hogg, R. V. & Craig, A. T. (1978). *Introduction to mathematical statistics*. New York, NY: Macmillan.

Klahr, D. & Nigam, M. (2004). The equivalence of learning paths in early science instruction. *Psychological Science*, *15*, 661–667. doi:10.1111/j.0956-7976.2004.00737.x

Kraemer, H. C. (2005). A simple effect size indicator for two-group comparisons? A comment on $r_{\text{equivalent}}$. *Psychological Methods*, *10*, 413–419. doi:10.1037/1082-989X.10.4.413

Kulik, C.-L., Kulik, J. A. & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research*, *60*, 265–299. doi:10.2307/1170612

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. doi:10.2307/2529310

Landry, S. H., Crawford, A., Assel, M., Gunnewig, S. B. & Swank, P. (2004). *Teacher Behavior Rating Scale (TBRS)*. San Antonio, TX: Unpublished research instrument, Center for Improving the Readiness of Children for Learning and Education.

Lee, V. & Burkam, D. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.

Logan, K. R., Bakeman, R. & Keefe, E. B. (1997). Effects of instructional variables on engaged behavior of students with disabilities in general education classrooms. *Exceptional Children*, *63*, 481–498.

Maas, C. J. M. & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, *46*, 427–440. doi:10.1016/j.csda.2003.08.006

Makar, B. W. (1995). *Primary phonics*. Cambridge, MA: Educators Publishing Service.

McCoy, K. M. & Pany, D. (1986). Summary and analysis of oral reading corrective feedback research. *Reading Teacher*, *39*, 548–554.

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. doi:10.1037//1082-989X.1.1.30

Meltzoff, A. N., Kulh, P. K., Movellan, J. & Sejnowski, T. J. (2009, July 17). Foundations for a new science of learning. *Science*, *325*, 284–288. doi:10.1126/science.1175626

Meyer, W. (1982). Indirect communication about perceived ability estimates. *Journal of Educational Psychology*, *74*, 888–897. doi:10.1037//0022-0663.74.6.888

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*, 376–390. doi:10.1037//0033-2909.86.2.376

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Murray, D. M., Hannan, P. J., Pals, S. P., McCowen, R. G., Baker, W. L. & Blitstein, J. L. (2006). A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Statistics in Medicine*, *25*, 378–388. doi:10.1002/sim.2233

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Pub. No. 00-4769). Washington, DC: U.S. Government Printing Office, Department of Health & Human Services, National Institute of Health. Retrieved from http://www.nationalreadingpanel.org.

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Pany, D. & McCoy, K. M. (1988). Effects of corrective feedback on word accuracy and reading comprehension of readers with learning disabilities. *Journal of Learning Disabilities*, *21*, 546–550. doi:10.1177/002221948802100905

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

Pianta, R. C., La Paro, K. M. & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual K-3*. Baltimore, MD: Paul H. Brookes.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*, 206–230. doi:10.3102/0002831207312905

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rosenshine, B. (1995). Advances in research on instruction. *Journal of Educational Research*, *88*, 262–268. doi:10.1080/00220671.1995.9941309

Rosenthal, R. & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). San Francisco, CA: McGraw-Hill.

Rosenthal, R. & Rubin, D. B. (2003). $r_{\text{equivalent}}$: A simple effect size indicator. *Psychological Methods*, *8*, 492–496. doi:10.1037/1082-989X.8.4.492

Rowan, B. & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher*, *38*, 120–131. doi:10.3102/0013189X09332375

Salvia, J. & Yssledyke, J. E. (1998). *Assessment* (7th ed.). New York, NY: Houghton Mifflin.

Samuels, S. J. (1997). The method of repeated readings. *Reading Teacher*, *32*, 403–408.

SAS Institute. (2005). *SAS OnlineDoc® 9.1.3: SAS/STAT 9.1 user's guide*. Cary, NC: Author.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi:10.1037//1082-989X.7.2.147

Share, D. L. (2008). Orthographic learning, phonological recoding, and self-teaching. *Advances in Child Development and Behavior*, *36*, 31–82. doi:10.1016/S0065-2407(08)00002-5

Shaywitz, S., Morris, R. & Shaywitz, B. (2008). The education of dyslexic children from childhood to young adulthood. *The Annual Review of Psychology*, *59*, 451–475. doi:10.1146/annurev.psych.59.103006.093633

Shoukri, M. M., Asyali, M. H. & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research*, *13*, 251–271. doi:10.1191/0962280204sm365ra

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. doi:10.1037//0033-2909.86.2.420

Simos, P. G., Fletcher, J. M., Bergman, E., Breier, J. I., Foorman, B. R., Castillo, E. M. & Papanicolaou, A. C. (2002). Dyslexia-specific brain activation profile becomes normal following successful remedial training. *Neurology*, *58*, 1203–1213.

Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Slocum, T. A., Street, E. M. & Gilberts, G. (1995). A review of research and theory on the relation between oral reading rate and reading comprehension. *Journal of Behavioral Education*, *5*, 377–398. doi:10.1007/BF02114539

Smith, D. (1979). The Improvement of children's oral reading through the use of teacher modeling. *Journal of Learning Disabilities*, *12*(3), 172–175.

Smith, M. W., Dickinson, D. K., Sangeorge, A. & Anastasopoulos, A. (2002). *Early language and literacy classroom observation toolkit* (Research ed.). Baltimore, MD: Brookes.

Smolkowski, K., Danaher, B. G., Seeley, J. R., Kosty, D. B. & Severson, H. H. (2010). Modeling missing binary outcome data in a successful web-based smokeless tobacco cessation program. *Addiction*, *105*, 1005–1015. doi:10.1111/j.1360-0443.2009.02896.x

Snijders, T. & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Snow, C. E., Burns, S. M. & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy of Sciences, National Research Council, Commission on Behavioral and Social Sciences and Education.

Stallings, P., Robbins, P., Presbrey, L. & Scott, J. (1986). Effects of instruction based on the Madeline Hunter model on students' achievement: Findings from a follow-through project. *Elementary School Journal*, *86*, 571–587. doi:10.1086/461468

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY: Guilford.

Stevens, C., Fanning, J., Coch, D., Sanders, L. & Neville, H. (2008). Neural mechanisms of selective auditory attention are enhanced by computerized training: Electrophysiological evidence from language-impaired and typically developing children. *Brain Research*, *1205*, 55–69. doi:10.1016/j.brainres.2007.10.108

Sutherland, K. S., Alder, N. & Gunter, P. L. (2003). The effect of varying rates of opportunities to respond to academic requests on the classroom behavior of students with EBD. *Journal of Emotional and Behavioral Disorders*, *11*, 239–248. doi:10.1177/10634266030110040501

Sutherland, K. S. & Wehby, J. H. (2001). Exploring the relation between increased opportunities to respond to academic requests and the academic behavioral outcomes of students with emotional and behavioral disorders: A review. *Remedial and Special Education*, *22*, 113–121. doi:10.1177/074193250102200205

Swanson, H. L. & O'Connor, R. (2009). The role of working memory and fluency practice on the reading comprehension of students who are dysfluent readers. *Journal of Learning Disabilities*, *42*, 548–575. doi:10.1177/0022219409338742

Temple, E., Deutsch, G., Poldrack, R., Miller, S., Tallal, P., Merzenich, M. & Gabrieli, J. (2003). Neural deficits in children with dyslexia ameliorated by behavioral remediation: Evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 2860–2865. doi:10.1073/pnas.0030098100

Vadasy, P. F., Sanders, E. A. & Peyton, J. A. (2005). Relative effectiveness of reading practice or word-level instruction in supplemental tutoring: How text matters. *Journal of Learning Disabilities*, *38*, 364–380. doi:10.1177/00222194050380041401

van Belle, G. (2002). *Statistical rules of thumb*. New York, NY: John Wiley & Sons.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R. & Denckla, M. B. (1996). Cognitive profiles of difficult to remediate and readily remediated poor readers: Toward distinguishing between constitutionally and experientially based causes of reading disability. *Journal of Educational Psychology*, *88*, 601–638. doi:10.1037/0022-0663.88.4.601

Wagner, R. K., Torgesen, J. K. & Rashotte, C. A. (1999). *CTOPP, Comprehensive Test of Phonological Processing*. Austin, TX: PRO-ED.

Walter, S., Eliasziw, M. & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, *17*, 101–110, doi:10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.3.CO;2-5.

Woodcock, R. W. (1998). *Woodcock Reading Mastery Tests (Revised). NU examiner's manual*. Circle Pines, MN: American Guidance Service.