Assessing the Relationship Between Observed Teaching Practice and Reading Growth in First Grade English Learners: A Validation Study

Scott K. Baker

Pacific Institutes for Research

Russell Gersten

Instructional Research Group

Diane Haager

California State University at Los Angeles

Mary Dingle

California State University at Sonoma

Claude Goldenberg

California State University at Long Beach

Abstract

Validation of a classroom observation measure for use with English Learners (Els) in Grade 1 is the focus of this study. Fourteen teachers were observed during reading and language arts instruction with an instrument used to generate overall ratings of instructional quality on a number of dimensions. In these classrooms, the reading performance of all ELs, as well as a sample of native English speakers, was assessed at the beginning and end of the school year to derive measures of reading growth over the course of the year. Technical characteristics of the observation measure and the reading growth of ELs are described. The relationship between classroom instruction and ELs' reading growth is interpreted within the context of a framework of measurement validity developed by Messick (1989), which proposes an integrated conception of validity. This framework is used to discuss data analysis and interpretation, and what implications and consequences these interpretations might have on instructional practices and professional development in early reading with ELs.

In this article, we describe a small scale, but intensive, classroom observation study. We focus as much on the validation of the observation measure as we do on the findings from the study. We do this because developing valid and reliable observational tools is essential for (a) providing teachers with meaningful feedback on their classroom practice, (b) understanding patterns of implementation that can direct professional development efforts, and (c) helping to better understand factors that accelerate reading achievement.

Our method of validation follows the framework developed by Messick (1989, 1995) and extended by Gersten, Keating, and Irvin (1995) and Gersten and Baker (2002). In this view, (a) construct validity is not easily separated from other types of validity, and (b) discussions of a valid assessment approach should include not only technical concerns, but should also explore the meaning of the data and corresponding actions that are taken on the basis the interpretations of the assessment data. In other words, discussions of the validity of a measure intentionally integrate correlational and/or factor analytic findings with discussions of underlying theories and instructional implications.

Background of the Study

The present study began in 1999, when we were approached by the directors of two professional development centers in California to create a research study that would provide them with knowledge regarding how to teach English Learners (ELs) to read in English. California had recently revamped its Reading and Language Arts Framework (California Department of Education, 1998, 1999) to reflect the findings of a report released by the National Research Council (Snow, Burns, & Griffin, 1998) on beginning reading. The purpose of this framework was to base reading instruction on scientific research on beginning reading. This was a dramatic and radical shift from the literature-based framework of a decade earlier. The Framework was a precursor to the principles and guidelines for initial reading instruction that would be codified, at the Federal level, under Reading First four years later (No Child Left Behind Act, 2001). The state guidelines clearly indicated that this framework was intended for all students, including ELs.

At the same time, schools in the state had just begun to implement Proposition 227, a law that significantly increased the number of ELs who were taught predominantly in English. The law required schools to teach ELs to learn to read in English unless their parents explicitly requested native language reading instruction. As a result of this legislation, the percentage of ELs who learned to read in English increased from 62% to 82% in three years (Merickel et al., 2003). The current study was conducted during the second year of implementation of both of these policies.

The NRC Report concluded that difficulties in phonological processing are causally linked to the majority of reading difficulties children experience (Adams, 1990; Stanovich, 1986), and can be prevented by explicit instruction in phonological processing that should begin as early as kindergarten (Snow et al., 1998). The California Framework stresses that successful instruction in phonological processing and phonics needs to be systematic, explicit, and intense in order to provide the greatest benefit to the greatest number of children (Gersten et. al., in press) The Framework also stresses the importance of intense, systematic instruction in comprehension, vocabulary, and reading fluency.

The Search for Exemplary Classrooms and Exemplary Practices

Initially, we planned to use the California state achievement database to identify schools and classrooms where the reading performance of ELs was higher than one would predict by demographic variables. We would then observe instruction in these schools and classrooms to determine the types of instructional practices being used. However, at that point in time, only a small percentage of ELs were being tested in any language in the primary grades. Thus, we were unable to obtain reliable estimates of reading achievement from the California database. We therefore decided to conduct our own reading assessments with students individually in classrooms at the beginning and end of the year to measure growth in reading.

Our goal was to develop our own classroom observational measure, and use these data to (a) describe overall patterns of practice in urban schools with large numbers of ELs in classrooms that were implementing beginning reading instruction according to the California Reading and Language Arts Framework, (b) obtain descriptive information on how teachers tailored reading instruction for ELs, and, most importantly, (c) determine whether we could link observed classroom practice to growth in reading.

Underlying conceptual framework

We followed Messick's (1989) framework by articulating a theory to guide instrument development. We hypothesized that the instructional practices articulated in the state Framework would be linked to accelerated achievement growth in reading for ELs. In this sense, our thinking paralleled Neufeld and Fitzgerald (2001) who, based on their qualitative research, concluded, "there is little evidence to support the need for a special vision of second-language reading instruction" (p.520). However, based on our review of the literature and own earlier observational research (e.g., Gersten, 1996; Gersten & Baker, 2000a, 2000b) we did think that certain adjustments and modulations might be necessary for ELs. In particular, we felt that early reading instruction for ELs should include a strong vocabulary component and provide students with frequent opportunities to practice talking about concepts learned in English. Finally, we wanted to ensure that we developed a measure that would capture the systematic, intensive, highly interactive style of reading instruction recommended by the National Research Council (Snow et al., 1998) report and the state Framework.

Purpose of the Present Study

Beginning in 1999, we developed, piloted and validated an observational system for use in first-grade classrooms where a majority of students (and often the whole class) were ELs. In earlier articles (Gersten, Baker, Haager & Graves, in press; Graves, Gersten & Haager, 2004), we described the development of the observational system.

This article focuses on the validation of the measure for use in first-grade classrooms where teachers are working with ELs. We focus, in particular, on criterion-related validity, that is, how well this measure of observed reading instruction predicts reading growth for ELs. We also discuss the validity of the measure in terms of the "four faces of validity" framework introduced by Messick (1995) and extended into the area of instructional practice by Gersten and Baker, (2002), and Gersten, Keating, and Irvin (1995). We use the Messick framework to explore issues related to use of an observational measure and meanings of the scores we obtained.

The path chosen to develop this measure is somewhat unusual. We decided to use a relatively high inference rating scale as opposed to the relatively low inference measures commonly used for reading instruction in the elementary grades (e.g. Taylor, Pearson, Peterson & Rodriguez, 2005; Connor, Morrison, & Katch, 2004; Foorman & Schatschneider, 2003). We made this decision largely because Likert type rating scales often correlate with achievement more highly than more objective measures of rates and frequencies of specific types of instructional activities (Gersten, Carnine, Zoref & Cronin, 1986). Before describing the design and details of the study, we provide a brief overview of relevant research and describe the context of the study.

Relevant Research

In this section we briefly review relevant research on teaching reading in a second language, much of which was conducted at approximately the same time that we were conducting this study. Recent longitudinal studies of ELs in the U. S. and Canada have demonstrated that teaching of phonological skills prior to – or at the onset of – formal reading instruction enhances reading skills for ELs. In fact, under these conditions, ELs may actually outperform peers in word-reading and decoding tasks (Lesaux & Siegel, 2003). Both Chiappe, Siegel, and Wade-Wooley (2002) and Geva, Yaghoub-Zadeh, and Schuster (2000) found that oral language proficiency in English is not predictive of how well or how quickly ELs will learn to read in English. One implication of this finding is that it may not be necessary for ELs to attain a certain level of oral language proficiency before they can learn to read in English, provided they have attended school on a regular basis and have been exposed to appropriate and systematic literacy instruction in English.

In fact, Vaughn et al. (this issue) posit that “there is evidence that development of [English] language proficiency….can be enhanced through reading instruction.” They note the converging evidence from an array of studies that English language reading proficiency often outpaces students’ oral language development (e.g., Fitzgerald & Noblit, 1999; Linan-Thompson, Vaughn, Hickman-Davis & Kouzekanani, 2003). Thus, they argue for a merger of English language reading instruction with other types of English language development instruction. The results of their experimental study would seem to support our working hypothesis.

However, it is important to discriminate between factors that predict how likely a child learns how to read in the early grades and with factors that predict long term reading success. Long term studies of reading development by Catts, Hogan, and Adlof (2005) found that although listening comprehension predicts a relatively small amount of unique variance in reading scores for second graders, by fourth grade, it uniquely predicts 21% of the variance, and by eighth grade, it uniquely predicts 36%. They note similar findings for ELs reported by Hoover and Gough (1990). Implications for these findings seems to be that a solid reading program for ELs in the primary grades should include a serious, systematic listening comprehension component, which should include comprehension of narrative and expository passages as well vocabulary development and understanding of English language grammar and syntax. Thus, it appears that the framework we used to assess important qualities of first grade reading instruction is consonant with contemporary research.

Development and Validation of the Observation Instrument

The English Language Learner Classroom Observation Instrument (Baker, Gersten, Goldenberg, Graves, & Haager, 1999; Gersten et al., in press; Haager et al., 2003) was used for all classroom observations (see Appendix A). Items were adapted from many sources. However, the core items were adaptations of Likert-scale items used by Stanovich & Jordan (1998) in their study of reading instruction. The Stanovich and Jordan measure was based on Englert's (1984) synthesis of research on effective reading instruction, integrating both the observational research of the 1980s (e.g., Brophy & Good, 1986) and more recent cognitive research on learning. We also used California's Reading Language Arts Framework as a source for observation items. Tikunoff et al. (1991) served as a major source for items on sheltered instruction. We adjusted the wording of items and sometimes the nature of the items to fit our target: first grade reading instruction.

We field-tested the instrument in 20 classrooms. We also made revisions based on extensive pilot testing (See Gersten, et al., in press, Haager et al., 2003, Graves, Gersten & Haager, 2004 for further detail on the development process). The primary sources for the items reflected our theory that, in general, techniques that are effective for students from high poverty backgrounds will be effective for ELs, but that use of sheltered instructional techniques will enhance comprehensibility and thus effectiveness.

Scoring and Development of Subscales and Reliability

Each item was rated on a Likert scale at the end of the day's reading lesson, which typically lasted between two and three hours (Gersten, et al., in press; Haager et al., 2003). During the reading lesson, the observer took detailed notes relating to the content of the items (e.g., examples of explicit modeling or ensuring all students participate in small group instruction). These notes were then used to guide the observer in completing the rating.

For each item, quality of instruction was rated on a 1-4 scale with 4 being the highest overall quality and 1 being the lowest overall quality. We also allowed mid-point ratings between 1 and 4 (i.e., 1.5, 2.5, and 3.5) to detect fine shades of differences in quality. These midpoints functionally created seven choices for each item. For analysis, we used the 4-point scale with the three midpoints.

Internal consistency for the measure was extremely high, with an overall coefficient alpha of .97. The six subscales in Table 1 demonstrated adequate internal consistency. Four of the six subscales had coefficient alphas above .80, one subscale was .78, and one was .65. We discuss these further in the results section.

Estimate of Criterion-Related Validity

Because we did not collect student reading data in the fall of the pilot testing year (i.e., Year 1), our criterion measure was growth from winter to spring in both oral reading fluency and comprehension. Without a fall assessment, these winter-spring measurement points provided a rough approximation of reading growth throughout the full year. However, even with this limitation, each of the subscales correlated moderately well with student growth in reading from the winter to spring of first grade (median correlation of .60).

Problems with Inter-rater reliability

Inter-rater reliability was lower than anticipated for the pilot testing phase. With a similar, but simpler observation instrument, Stanovich and Jordan (1998) reported an exact agreement rate among observers of 78%, which was higher than the first year data we collected as reported in Gersten et. al (in press). Difficulty establishing high reliability estimates may have been caused by the length and complexity of the instrument, the nature of the rating procedure, or the limited training time observers had to learn to use the instrument in a common way. Most likely it was a combination of these three factors.

Revision of the Observational Measure for the Current Study

To improve inter-observer reliability, we reduced the length of the instrument in year 2. The high internal consistency suggested this was a reasonable course of action. We deleted items that demonstrated either low item-to-total correlations (below .3) or low inter-observer reliability. We also removed items that seemed redundant, or items that observers said were unclear. The final instrument contained 33 items.

Our objectives were to (a) determine whether the observational measure was a valid means of estimating a class' growth in reading over the course of first grade, (b) determine whether this highly inferential type of instrument was a reliable and feasible measure, and (c) determine whether the subscales represented reliable, and important facets of instruction.

We wanted to focus on overall patterns of practice in urban schools with large numbers of ELs in classrooms that were implementing beginning reading instruction according to the California Reading and Language Arts Framework. We were interested in particular in how teachers tailored reading instruction for ELs, and whether these instructional variations could be to growth in reading.

Method

Participants

Teachers

Fourteen first-grade teachers participated in the study. These teachers were selected from seven schools in four California school districts. In these schools, instruction followed California's adopted Reading and Language Arts Framework, which required use of research-based practices for a minimum of two-and-a-half hours per day. The teachers met two additional criteria: (a) they possessed at least three years of teaching experience in the primary grades, and (b) they taught in a class where at least half the students were Els. Reading instruction was primarily in English, although the students' native language may have been used on occasion to explain concepts or give specific instructions.

Approximately two thirds of the teachers used a core reading program that was based on contemporary research and attempted to systematically develop phonological processing and decoding skills. The other teachers used a more literature-based approach, relying mainly on trade books for their core instruction. Yet these teachers also taught in accordance with California's Framework by emphasizing systematic phonemic awareness, decoding, vocabulary, and comprehension through the use of teacher developed mini-lessons and the periodic use of basal or supplemental reading materials.

Eleven teachers reported data on years of teaching experience and 10 on their ethnicity. Of this group, almost half (5 of 11) had between 3 and 5 years experience, 3 had between 6 and 15, and 3 had more than 20 years of experience. Twenty percent of the teachers were Hispanic, 10% Asian-Pacific Islander, 20% African-American and half were Caucasian.

Students

In the 14 target classrooms, a total of 194 first grade students were assessed at both pretest and posttest. Among these students, 14 different primary languages were spoken; 110 of these students spoke Spanish as their primary language and 53 spoke English. Twelve additional primary languages were spoken among the remaining 31 students. For purposes of data analysis, these 31 students were combined into one group. All 194 students were eligible for free or reduced lunch rates.

Student Measures

Reading performance was assessed using the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Kaminski & Good, 1996; Good & Kaminski, 2002). The DIBELS measures are a series of one-minute tasks that measure constructs related to phonological awareness, alphabetic understanding, and reading fluency. A constructed response reading comprehension measure adapted from the California Reading Results Reading Comprehension Assessment (California Reading and Literature Project, 1999) was administered to students at the end of first grade. A brief description of each measure, in terms of the underlying construct, follows.

Phonological Awareness

Phonemic segmentation fluency (Kaminski & Good, 1996; Good & Kaminski, 2002). Examiners orally presented 2, 3, and 4-phoneme words to students. Students responded by saying the individual phonemes in the word. They received one point for each correct phoneme they produced (i.e., zero to four points per word). Alternate-form reliability of the measure is reported at .88, and predictive validity over one year with reading measures ranged from .73 to .91 (Kaminski & Good, 1996). The task was modeled and practiced prior to administration.

Letter naming fluency (Kaminski & Good, 1996; Good & Kaminski, 2002). Students were presented with randomly ordered upper and lower case letters arranged in rows on an 8.5 by 11-inch piece of paper and asked to name as many letters as possible in one minute. Reliability of the measure has been reported at .93 by Kaminski and Good (1996); one-year predictive validity coefficients with reading criterion measures ranged from .72 to .98.

Nonsense word fluency (Good, & Kaminski, 2002; Good, Gruba & Kaminski, 2003). Nonsense Word Fluency measures a student’s proficiency at utilizing the alphabetic principle. Students were presented with a series of VC and CVC pseudo words (e.g., et, zeb) arranged in rows on an 8.5 by 11-inch piece of paper. They were asked to say the sounds of the letters, or read the “words.” The number of correct sounds produced in one minute, either in isolation or within the correctly read nonsense word, was determined. Good and his colleagues reported a one-month alternate form reliability of .83 for students in the middle of first grade. Concurrent criterion-related validity with

Woodcock-Johnson Psycho-Educational Battery-Revised Readiness Subscale score was reported at .59 in February of first grade. Predictive validity with Oral Reading Fluency was reported at the end of first and second grade as .82 and .60, respectively.

Reading Fluency

Oral Reading Fluency (Shinn, 1998). Each student read aloud a story written at a first-grade level and the number of words the student read correctly in one minute provided the index used in data analysis. Estimates of the internal consistency, test-retest, and inter-scorer reliability for Oral Reading Fluency have ranged from .89 to .99. Correlations with other measures of reading, including measures of decoding and comprehension, have ranged from .73 to .91 (Shinn, Tindal, & Stein, 1988). Correlations between Oral Reading Fluency and standardized measures of reading comprehension are typically above .80 (Marston, 1989). Baker and Good (1995) examined the technical characteristics of Oral Reading Fluency with ELs and native English speakers in grade 2. The measure was as reliable, and worked as well as an index of comprehension, for Els as for native English speakers. Interestingly, Oral Reading Fluency was a better measure of reading progress for Els than it was for native English speakers. Baker and Good's findings suggest that Oral Reading Fluency measures the same reading construct for both Els and native English speakers.

Constructed Response Reading Comprehension Assessment

We used an adaptation of a reading comprehension measure used in the California Reading and Literacy Project (California Reading and Literature Project, 1999) to assess comprehension. This project provided a sample of assessment tasks that the state expected proficient readers to be able to accomplish independently. On the constructed reading comprehension tasks, students were asked to read a short story and then write answers to five questions about the story (the story and the questions are included in Appendix B). Students were given 30 minutes to read the story and complete the questions. After surveying the range of student responses to each question, scoring guidelines were established. The number of points possible for each question ranged from 1 to 3 depending on the range of student responses. Reliability on the 25% of protocols scored by two project staff members ranged from .87 to .95.

Procedures

Observation Procedures

A total of 31 observations were conducted in the classrooms of the 14 participating teachers. Each observation took place during an entire Reading/Language Arts instructional period. Five researchers, selected on the basis of their experience with classroom observations, English Learners, and reading instruction, conducted these observations. In 10 of the 14 classrooms, observations were conducted by at least two members of the observation team in an effort to control for individual observer effects and to explore the issue of reliability in depth.

Most of the observations occurred between the fourth and sixth months of the school year. We considered this the best time to observe because with the exception of phonemic awareness we expected teacher to be implementing the components of reading instruction specified in the measure.

Student Assessment Procedures

Students were assessed in the beginning and end of the school year. Initial assessments occurred between one and two months after the start of school, and final assessments occurred between one and two months prior to the end of the school year. All measures, except the open-ended reading comprehension measure, were administered individually to children in a quiet place in or near their classroom. All test administrators received training in the administration and scoring of each measure. The open-ended reading comprehension measure was group administered. One member of the project staff administered and proctored the test. Students received as much time as necessary to complete the test.

Results

In the first section, we present data on the reliability and validity of the observational instrument. We present data on the internal consistency of the total measure and each of the empirically derived subscales as well as the item-by-item inter-observer reliability. We use the subscale reliabilities to discuss the construct validity of the measure. In particular, we examine what Messick (1988, 1989) calls the *evidential basis* for interpretation of the empirically derived subscales as important aspects of effective instruction. These data include the stability of the subscale structure over time, the extent to which the subscales correlate with each other (and thus may in fact be redundant with each other). In the final section we discuss the extent to which the measures of observed teaching practice predicted student reading outcomes (i.e., criterion-related validity) or evidence of the potential for valid use of the observational system.

Reliability of the Observational Instrument

Internal consistency of the measure. The internal consistency of the observation instrument remained high in Year 2, with a coefficient alpha of .95. We were interested in examining whether the subscales developed in Year 1 remained reliable. As noted by Messick (1989), subscale reliability is also a measure of construct validity.

Subscale internal consistency. The subscales were developed empirically using Year 1 data. We thought it would be better to create subscales empirically rather than on an *a priori* basis because even though the items came from numerous sources, we wanted to create scales that cut across the different research traditions (observational research on reading, research on sheltered instruction, experimental research on reading). Although factor analysis typically requires a large sample of teachers and classrooms, we felt the use of this procedure in an exploratory way to generate empirically derived subscales was appropriate.

Confirmatory factor analysis would be an ideal means for determining the validity of the factor structure (i.e., subscales) generated in Year 1, the small sample size precluded its use. We therefore began by determining if the same subscale structure remained reliable with the new data set.

In Table 1 we present the coefficient alpha coefficients for each subscale and present the items that comprise each subscale However, with one exception, the subscales are reliable with coefficient alphas ranging from .65 to .91 and a median reliability of 86.5. The only subscale that appears problematic is *Interactive Teaching.*

Correlations Between the Subscales: Exploring Underlying Constructs

In Table 2, we present the correlations between the six empirically derived subscales. The highest correlation is between Sheltered English Instruction and Vocabulary Instruction; both involve aspects of language instruction that may be particularly salient for Els. These data suggest that both subscales may reflect one dimension of teaching practice. In future uses of the measure, we would consider them one subscale. Note that these subscales do not demonstrate significant correlations with Phonics/Decoding, as one might expect. As we observed, we noted that some teachers were strong in language and vocabulary development, but not necessarily in systematic instruction in decoding. Others reflected an opposite pattern.

The correlation between Subscale 2 (Quality of Instruction for Low Performers) and Subscale 3 (Sheltered Instruction) was also not significant. The pattern of correlations suggest a dimension relating to language-sensitive instruction, a term developed by Chamot and O'Malley (1996). Correlations between the Explicit Teaching and Attention to Low Performers subscales are high. They also seem to correlate moderately well with the interactive teaching factor. In future use of the measure, we would consider merging the four items of the interactive teaching subscale with a larger subscale.

Inter-observer agreement. We conducted inter-rater reliability in 8 of the 14 classrooms (i.e., 56% of the classrooms involved in the study). Reliability observations had two observers in the classroom at the same time or within two instructional days of each other. It is important to remember that each observation lasted for two-and-a-half hours and the goal was to obtain a rating of a teachers' typical instructional style and qualities. Thus, we saw benefits in seeing whether scores concurred when made within two days of each other.

Reliability estimates were calculated by summing the number of rating agreements across the entire instrument and dividing that total by the number of agreements plus disagreements. Since this was a high inference rating scale, we considered either an exact match or a half-point difference on an item as an agreement, and a difference of 1 whole point or more on the 4-point scale as a disagreement. Across the 8 reliability observations, the median percent agreement was 73% (mean reliability = 67%), ranging from 39% to 97%. This calculation of item-by-item agreement is a stringent criterion for use in a rating scale. Often reliabilities are conducted only on scores for an entire scale.

Inter-observer agreement on an item-by-item level was respectable, but lower than the .78 obtained by Stanovich and Jordan (1998). Across all of the items on the instrument, we noted that the more highly inferential items tended to demonstrate lower reliability. For example, the item “Selects and incorporates students responses, ideas, examples and experiences into lesson” or “Structures opportunities to speak English” are more difficult to operationalize than items such as “Secures and maintains student attention during lesson” or ‘Makes relationships overt” or “Provides prompts.” On the more behavior items, observers demonstrated higher rates of agreement, often reaching 100%. In addition, items in the subscales that deal with language (sheltered instruction and vocabulary development), which clearly are more difficult to define operationally, tended to have lower agreement than those items linked to management and instructional presentation.

Criterion Related Validity: Evidence of the Utility of the Observational Measure

Before discussing the relationship between scores on the observational measure and growth in student reading, we present a brief overview of student performance on the reading and reading related measures.

Student Performance

Performance at pretest and posttest of the 194 students in the 14 target classrooms is presented in Table 3.

Pretest

There is little consistency in the pattern of performance of the three linguistic groups on measures likely to predict subsequent reading proficiency. On Letter Naming Fluency (in English), the native Spanish-speaking students were more than .5 standard deviations below the other groups. Yet, on the phonological processing measure, the students with other primary languages were almost a full standard deviation below both the native English speakers and the Spanish speakers. Performance on the Nonsense Word Fluency was more or less identical between the three groups. Although the pretest differences were relatively small, it may indicate that the Other Primary Languages group was at a more advanced stage of reading than the other two groups, and their emerging knowledge of letter-sound correspondence may have interfered with their performance on a pure measure of phonological processing.

Posttest.

The highest performing group on the two reading measures is the Other Primary Language group. They score approximately .38 standard deviation units higher than the native English speakers and .84 standard deviation units higher than the native Spanish speakers. On average, the Oral Reading Fluency score for the Other Primary Language group approaches a reasonable benchmark for English speaking children.

The pattern is virtually identical for Reading Comprehension, with Spanish speakers .48 standard deviation units below the English speakers and approximately .72 standard deviation units below the Speakers of Other Primary Languages. Scores are relatively similar, in contrast, on the Nonsense Word Fluency measure. This measure assesses only the ability to quickly read simple CVC and CVCe nonwords, and may demonstrate ceiling effects for the stronger readers.

The contrast between the Spanish speakers and Speakers of Other Primary Languages samples is a finding worthy of further attention. The Other Primary Languages group, on average, is doing well on measures of reading performance in first grade and may be benefiting the most of the three groups from first grade instruction.

Relationship Between Classroom Instruction And Student Growth In Reading

We analyzed these data using a two-level nested model by applying hierarchical linear modeling (HLM) (Raudenbush, Bryk & Richard, 2004). In this analysis the individual student served as the unit of analysis at level one and the class/teacher as the unit of analysis at level two. This type of model allows level one and level two covariates to interact as predictors of achievement gains. For the HLM analyses, we chose two subscales that seemed to represent relatively unrelated dimensions of teaching: *Sheltered Instruction Techniques* and *Explicit Teaching*.

We were aware, however, of the low statistical power when the class was used as the level of analysis, since the sample size was only 14 at this level. Therefore, we also conducted multiple regression analyses using the student as the unit of analysis. We also conducted correlational analyses using the teacher as the unit of analysis, examining the relations between overall measures of classroom instruction with mean rates of growth on measures of reading performance.

Choice of Unit of Analysis for Regressions

There are several advantages when using the class as the unit of analysis. The observation measure focused on teachers working with their entire classes (even if they sometimes broke the class into small groups, in which case we circulated around the room but maintained our primary focus on the teacher). Thus, the predictor variable is the observed quality of instruction provided to the entire class. Teachers may have provided higher quality instruction to subgroups of students but our observation system was not sensitive to these types of instructional variations because we did not focus at the individual student level. Consequently, in this system there is a logic to using the class as a unit of analysis. In addition, each observation is weighted equally, regardless of class size, differential attrition, or proportion of Els in the class.

Advantages of using the student as the unit of analysis include the ability to perform more complex statistical analyses since the sample size is much larger. Another advantage is that a class with 16 valid pretest and posttest measures receives more weight than a class with only 8, for example. The estimate of mean reading growth is appreciably more accurate for the class with 16 students than the class with 8 students.

In general, results were similar using HLM and more traditional regression analyses. Since the regression results are more familiar and thus easier for most readers to comprehend, we use this format to present the findings. However, two interesting findings from the HLM analyses seem worthy of mention. We discuss these first, followed by an analysis using the student as unit of analysis. We conclude with a brief overview of correlations using the class as the unit of analysis.

Results of HLM Analyses

The first important finding is that although neither of the two subscales selected (Explicit Teaching or Sheltered Instruction Techniques) was a significant predictor of posttest oral reading fluency at the classroom level (due to the small number of degrees of freedom at the classroom level (df = 11), both approached significance, *p* = . 09 for Subscale 1 (Explicit Teaching), and *p* = .11 for Subscale 3 (Sheltered Instruction Techniques). The most interesting finding is that there was an interaction between (a) Subscale 3 and (b) the contrast between Spanish-speaking and native English-speaking populations. This means that the use of sheltered techniques was differentially beneficial based on student group, being particularly useful for the Spanish speakers. This finding makes sense in that sheltered instruction is designed specifically to benefit ELs. In contrast, Subscale 1, Explicit Instruction, did not demonstrate a significant interaction. That is, it was similarly effective for both ELs and native English speakers.

In the traditional regression analysis, four pretest student performance variables were entered: Letter Naming Fluency and Phonemic Segmentation Fluency as “prereading measures,” and Nonsense Word Fluency as an “early reading measure.” Also entered were mean rating scores on Factor 1 (Explicit Teaching) and Subscale 3 (Sheltered English Techniques). The outcome measure was Oral Reading Fluency at posttest. In this analysis, a 3-variable model was significant and explained the most variance (F = 40.75, df = 3, 142, adjusted R square = .45). Letter Naming Fluency entered first (R square change = .34), then Sheltered English Techniques (R square change = .08), then Nonsense Word Fluency (R square change = .04).

With Reading Comprehension as the outcome variable, and the same predictor variables entered, the result indicated that another 3-variable model was statistically significant and accounted for 33% of the variance (F = 19.54, df = 3, 112; adjusted R square = .33). Letter Naming Fluency still entered first (R square change = .22), Explicit Teaching entered second (R square change = .09), and Phonemic Segmentation Fluency entered third (R square change = .03). It is noteworthy that in both regression analyses, instructional quality contributed to models explaining reading outcomes above and beyond reading performance at pretest. The effect sizes correspond to .33 for Explicit Instruction on the Reading Comprehension measure and .30 for Sheltered English Techniques on Oral Reading Fluency.

Using the Classroom As Unit Of Analysis

To determine the extent to which the observed appraisals of the quality of classroom reading instruction predicted reading growth over the course of the year, we first examined correlations between ratings of each subscale with a residualized growth score in reading calculated for each of the 14 classrooms. For these analyses, we created a *composite* reading score consisting of posttest performance on Oral Reading Fluency and the Reading Comprehension measure. The correlation between these two measures was .59, indicating that the two measures, though related, did in fact represent distinct facets of beginning reading. The composite, then, assesses not only the ability to read connected text in English accurately and fluently, but also the ability to understand stories read independently. Both seem critical outcomes for ELs and somewhat of an advance on prior research, which has often been limited to the reading of word lists.

On each of these two measures, we adjusted posttest performance based on pretest performance on the Letter Naming Fluency measure. This adjustment was made to control for initial differences in student ability between classrooms. We used Letter Naming Fluency because this measure proved to be the best predictor of Oral Reading Fluency for our sample, as well as a strong predictor of Reading Comprehension.

The relationship between scores on each of the subscales and the composite reading outcome measure is presented in Table 4. The correlations are statistically significant except the correlation between the Phonics/ Decoding Subscale and the Reading Composite score. This subscale only included two items and needs to be expanded or merged with other subscales. Correlations for each of the other four subscales, Explicit Instruction, Instruction Geared towards Low Performers, Sheltered Instruction, and Quality of Vocabulary Instruction, are moderately high in magnitude. The correlations with reading growth were equivalent to those found in the earlier study (Gersten et al., in press; Gersten & Baker, 2003).

The implication is that the observers’ perceptions of the quality of reading instruction corresponded moderately well to reading growth at the overall classroom level.

Furthermore, the various facets of instruction identified by five of the instructional subscales appear to be related to reading growth for ELs.

Discussion

In the discussion, we rely on a validity framework by Messick (1988, 1989) to explore the meaning of the findings and implications for the field. We begin by interpreting the findings in the context of what Messick labels *consequential validity.* We discuss the implications of the correlations and multiple regression analyses, which generally indicated there was a consistent relationship between the observations of reading instruction and students' growth in reading. Then, from the perspective of what Messick refers to as *value implications,* we discuss the data in relation to the body of knowledge on teaching students to read in a second language. In the section of Messick's framework that addresses *data interpretation*, we rely on a variation proposed by Gersten, Keating, and Irvin (1995) and Gersten and Baker (2002).

We also discuss how the observation instrument we developed for this study could be refined for future use in projects like Reading First, and in understanding instructional implementation in the context of other current State and Federal reading initiatives in education. In broad terms, we discuss both the implications of findings for improving the observation instrument, and for developing a better understanding of how to effectively teach Els to read in a second language.

Consequential Validity

The major finding of the study is that observers with a reasonable knowledge of beginning reading research were able to rate observed instructional practice in a valid fashion, that is, in a way that predicted classroom reading growth. The team of observers, despite different backgrounds related to second language acquisition, special education, and general education, and with somewhat different orientations to the teaching of reading, were able to rate instructional effectiveness in a way that was moderately related to reading growth over time.

One implication for this finding is that our attempt to fuse diverse bodies of knowledge into a single observation instrument may be on the right track. Items were generated from the knowledge base on scientific research on beginning reading, including observational research of Barbara Foorman and her colleagues, as well as work from an earlier era involving researchers such as Jane Stallings and Linda Anderson and Jere Brophy. The instrument also integrated more recent work in the cognitive strategy/constructivist field involving the research of Stanovich and Jordan (1998).

The observation instrument also included items that typified what researchers and administrators thought were exemplary practices for teaching ELs reading and other academic content in English (i.e., sheltered instruction). Our proposition was that these traditions needed to be merged since teachers are responsible for both reading instruction and ELD instruction. An overall instructional framework that integrated these two language-related teaching objectives would be optimal. Our sense is there is an untapped connection between these two bodies of research and the observation instrument represents an initial attempt to begin to merge heretofore diverse traditions (For another attempt at integration and synthesis, see Echevarria, Short, & Vogt, 2000).

To recapitulate specific findings, when the class was used as the unit of analysis, and the outcome measure was a composite of oral reading fluency and a researcher-developed comprehension measure (both adjusted for pretest performance on Letter Naming Fluency), significant and strong correlations were found not only for the total score on the observation instrument, but also for the following subscales, which represent various important facets of teaching: (a) Explicit Teaching, (b) Instruction Geared to Low Performers, (c) Interactive Teaching, (d) Sheltered Instruction, and (e) Vocabulary Development. Because Sheltered Instruction and Vocabulary Development are highly correlated, and in essence address language development, it may be best to think about them as representing one major facet of teaching. The Phonics / Decoding subscale was significantly related to growth in Year 1, but not Year 2. We remain convinced, however, the Phonics / Decoding subscale is an essential component of reading instruction for ELs, but that a more substantial and reliable scale is necessary to tap this specific dimension of teaching.

When the student was the unit of analysis, quality of instruction also emerged as a significant predictor above and beyond pretest performance. In order to understand the meaning of the data, it is important to understand that quality of instruction is not intended to level out individual differences, but to play a significant role in explaining the amount of variance above and beyond what a student brings to school in terms of natural aptitude and relevant background knowledge. For both Reading Comprehension and Oral Reading Fluency, observed ratings of quality of instruction explained a significant amount of the additional variance.

These two analysis objectives have strengths and weaknesses. Using the class as the unit of analysis is advantageous in that the observational ratings were based on our appraisal of how the teacher worked with the entire class, not with individual target students. Thus, the data are analyzed at what is probably the most appropriate level of analysis in terms of the target of observations during reading instruction. Drawbacks of this method, of course, are the small sample size, and consequent limits placed on the sophistication of data analyses that can be used.

When the student served as the unit of analysis, more sophisticated regression techniques can be used. However, some classes are weighted more heavily than others because weighting depends on the number of ELs who attended school at both pretest and posttest. Consequently, for example, a class with all ELs and no attrition could have triple the weight of a class with half ELs and moderate attrition. Another limitation of this technique is that the observation focuses on an entire class and will likely reflect differing degrees the quality of instruction received by any individual student.

Value Implications: Meanings of the Subscales

The fact that the initial factor analysis conducted with the pilot data from Year 1 was replicated in Year 2 suggests that there may well be a set of between four and six facets of teaching that are related to acceleration of reading growth for ELs in the early years of schooling. The fact that each of the subscales demonstrates a significant relationship to enhanced classroom reading performance suggests that each is important in some fashion.

Suggestions for Refining the Subscales for Future Use

The data suggest that the observation instrument might be refined for future use the he following ways:

It may be useful to collapse the vocabulary and sheltered instruction into one overall subscale. Because appropriate and frequent vocabulary instruction is so essential to effective instruction for ELs, it makes sense conceptually and in terms of classroom practices that vocabulary instruction would be strongly related to other strategies teachers would use in providing sheltered instruction for ELs.

Additional items should be added to the phonemic awareness/phonics/decoding subscale to increase its reliability and scope. Several of the items deleted in the first year of implementation might be useful for this purpose (See Appendix C). These include: (a) quality of spelling instruction, (b) opportunities provided for students to generate their own text, and (c) checks on students' oral reading fluency. By analyzing instruction related to these types of items, a revised scale would be richer in targeting an array of activities that involve phonemic awareness and knowledge of the alphabetic principle. This could well be a relatively precise gauge of the quality of the teachers' approach to teaching students how to read, as opposed to instruction more closely aligned with language and vocabulary development reflected in the Sheltered Instruction and Vocabulary Development subscales.

The interactive teaching subscale appears to be problematic in its current form, and may well be primarily an issue of reliability. Two of the four items have low inter-rater reliabilities and the scale may have too few items. This remains, in our view, an important construct in understanding quality of instruction.

Implications of the Constructs for Understanding Teaching: Consequential Validity

If additional analyses suggest that there are four to five related facets of effective reading instruction for students in a second language, we believe educators involved in professional development and teacher training might have a useful framework for conceptualizing relative strengths and weaknesses of a given teacher at a given point in time, which could provide a *focused, coherent* framework for professional development.

Current efforts at professional development often move with no particular rhyme or reason from vocabulary development to phonemic awareness, to the importance of fluency, to techniques for assisting struggling readers. If teachers and professional development personnel could share explicitly a dimension of teaching that would be the focus for sustained work, we believe professional development in beginning reading could begin to have a coherence (Garet, Porter, Desimone, Birman, & Yoon, 2001) that it currently lacks.

In a large national survey, Garet and colleagues found that teachers perceived professional development approaches as useful when they could see a coherence between the array of activities. Merely telling teachers that all the techniques are based on scientific research will not provide such coherence. Within each area – work on language and vocabulary development, strategies for interventions and support for low achieving students, phonemic awareness and phonics activities that are linked and reinforce each other – an array of related, coherent concepts and research based principles and suggestions for improving practice could be developed and discussed. Knowing that each plays a role in enhancing reading outcomes, even if we are far from establishing the exact nature of the interrelationships between them, could be a hinge for improving the quality of professional development.

The research reported in this article is most assuredly exploratory. Current, ongoing large scale research projects, such as the observational research conducted by Foorman and Saunders (this issue) will provide a deeper understanding of the extent to which our conceptual framework – the empirically derived facets of teaching – is replicated in other studies using larger samples and more complex methodologies. This understanding will also be refined, and perhaps challenged, by the findings of the National Literacy Panel established by the U. S. Department of Education to synthesize the research on second language reading.

We note, however, that recent research by Milanowski and Odden (2004) found that the median correlation of four well regarded "anchored" Likert rating scales on quality of reading instruction was .28; the lowest was .21 and the highest was .51. The fact that the items on the scale reported in this study correlate with reading growth in the range of .6 - .7 suggests that, given the vast body of reading research generally, we are moving towards a more refined understanding of the facets of high quality reading instruction for students learning to read in a second language specifically.

References

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

Baker, S., Gersten, R., Goldenberg, C., Graves, A., & Haager, D. (1999). *Reading and language arts classroom observation instrument for the early primary grades.* Unpublished Assessment.

Baker, S. K., & Good, R. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review, 24*, 561-578.

Brophy, J., & Good, T.L. (1986). Teacher behavior and student achievement. In M Witrock (Ed.), *The third handbook of research on teaching* (pp. 328-375). New York: Macmillan.

California Department of Education (1998). English and language arts content standards for grades K-12. Sacramento, CA: Author.

California Department of Education. (1999). Reading/language arts framework for California public schools, grades K-12. Sacramento, CA: Author.

California Reading and Literature Project. (1999). *Reading professional development institute focusing on results, K-3*. San Diego, CA: California Reading and Literature Project.

Chamot, A. U., & O'Malley, J. M. (1996). The cognitive academic language learning approach: A model for linguistically diverse classrooms, *Elementary School Journal* (Vol. 96, pp. 259-273).

Chiappe, P., Siegel, L., & Wade-Wooley, L. (2002). Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading, 6*, 369-400.

Connor, C., Morrison, F., & Katch, L. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, 8 (4), 305-336.

Echevarria, J., Vogt, M., & Short, D. J. (2000). *Making content comprehensible for English-language learners: The SIOP model*. Boston: Allyn and Bacon.

Englert, C. S. (1984). "Effective direct instruction practices in special education settings." *Remedial and Special Education (RASE) 5*(2): 38-47.

Foorman, B., & Schatschneider, C. (2003). Measurment of teaching practices during reading/language arts instruction and its relation to student achievement. In S. Vaughn & K. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning*. Baltimore, MD: Brooks.

Fitzgerald, J., & Noblit, G. W. (1999). About hopes, aspirations, and uncertainty: First-grade english language learners' emergent reading. Journal of Literacy Research, 31(2), 133-182.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915-945.

Gersten, R. (1996). Literacy instruction for language-minority students: The transition years, *Elementary School Journal* (Vol. 96, pp. 227-244).

Gersten, R., & Baker, S. (2000a). The professional knowledge base on instructional practices that support cognitive growth for English-language learners. In R. M. Gersten & E. P. Schiller (Eds.), *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues. The LEA series on special education and disability.* (pp. 31-79).

Gersten, R., & Baker, S. (2000b). What we know about effective instructional practices for English-language learners, *Exceptional Children* (Vol. 66, pp. 454-470).

Gersten, R., & Baker, S. (2002). The relevance of Messick's four faces for understanding the validity of high stakes assessments. In G. Tindal, & T. Haladena (Ed.), *Large-scale assessment programs for ALL students: Validity, technical adequacy, and implementation*. Mahwah, New Jersey: Lawrence Erlbaum.

Gersten, R., Baker, S., Haager, D., & Graves, A. (in press). Exploring the role of teacher quality in predicting reading outcomes for first grade English learners: An observational study. Remedial & Special Education.

Gersten, R., Carnine, D., Zoref, L., & Cronin, D. (1986). A multifaceted study of change in seven inner city schools. *Elementary School Journal*, *86*(3), 257-276.

Gersten, R., Keating T. J., & Irvin, L. K. (1995). The burden of proof: Validity as improvement of instructional practice. *Exceptional Children, 61,* 510-519.

Geva, E., Yaghoub-Zadeh, Z., & Schuster, B. (2000). Understanding differences in word recognition skills of ESL children. *Annals of Dyslexia, 50*, 123-154.

Good, R.H., Gruba, J., & Kaminski R.A. (2002). Best practices in using dynamic indicators of basic early literary skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 699-720). Bethesda, MD: NASP Publications.

Good, R. H., & Kaminski, R. A. (Eds.) (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu/.

Good, R. H., Kaminski, R. A., Smith, S., Simmons, D., Kame'enui, E., & Wallin, J. (2003). Reviewing outcomes: Using DIBELS to evaluate kindergarten curricula & interventions. In S. R. Vaughn & K. L. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 221 - 259). Baltimore: Brookes.

Graves, A., Gersten, R., & Haager, D. (2004). Literacy instruction in multiple language first grade classrooms: Linking student outcomes to observed instructional practice. Learning Disabilities Research & Practice, 19(4), 262-272.

Haager, D., Gersten, R., Baker, S., & Graves, A. (2003). The English-Language Learner Classroom Observation Instrument: Observations of beginning reading instruction in urban schools. In S. R. Vaughn & K. L. Briggs (Eds.), *Reading in the*

Classroom: Systems for Observing Teaching and Learning. Baltimore, MD: Paul Brookes Publishing.

Hoover, W.A. and Gough, P.B. (1990). The simple view of reading. Reading and Writing: An Interdisciplinary Journal, 2, 127-160.

Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*, 215-227.

Lesaux, N., & Siegel, L. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology, 39,* 1005-1019.

Linan-Thompson, S., Vaughn S., Hickman-Davis, P., & Kouzekanani, K. (2003). Effectiveness of supplemental reading instruction for second-grade english language learners with reading difficulties. The Elementary School Journal, 103(3), 221-238

Marston, D. (1989). Curriculum-based measurement: What is it and why do it? In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.

Merickel, A., Linquanti, R., Parrish, T., Perez, M., Eaton, M., & Esra, P. (2003). *Effects of the implementation of proposition 227 on the education of English learners, K-12: Year 3 report*. Washington, DC: American Institutes for Research and WestEd.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 33-46). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Education measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.

Milanowski, A., Kimball, S., & Odden, A. (2004). Teacher accountability measures and links to learning. Madison, WI: University of Wisconsin-Madison.

Neufeld, P., & Fitzgerald, J. (2001). Early English reading development: Latino English learners in the "low" reading group. *Research in the Teaching of English*, *36*, 64-109.

No Child Left Behind Act, 20 U.S.C. 6301 (2001).

Raudenbush, S.W. and Bryk, A.S. (2002). Hierarchical linear models (Second Edition). Thousand Oaks: Sage Publications, 482 pp.

Shinn, M.R. (Ed.). (1998). *Advanced applications of curriculum-based measurement*. New York: The Guilford Press.

Shinn, M. R., Tindal, G. A., & Stein, S. (1988). Curriculum-based measurement and the identification of mildly handicapped students: A research review. *Professional School Psychology, 3*, 69-85.

Snow, C. S., Burns, S. M., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Stanovich, K. E. (1986). Cognitive processes and the reading problems of learning-disabled children: Evaluating the assumption of specificity. In J. K. Torgesen & B. Y. L. Wong (Eds.), *Psychological and educational perspectives on learning disabilities* (pp. 87-131). Orlando: Academic Press**.**

Stanovich, P. J., & Jordan, A. (1998). Canadian teachers' and principals' beliefs about inclusive education as predictors of effective teaching in heterogeneous classrooms. *Elementary School Journal, 98,* 221-238.

Taylor, B., Pearson, P.D., Peterson, D., & Rodriguez, M. (2005). The CIERA school change framework: An evidence –based approach to professional development and school reading improvement. *Reading Research Quarterly,* 40, 40-68.

Tikunoff, W. J., Ward, B. A., van Broekhuizen, L. D., Romero, M., Castaneda, L. V., Lucas, T., et al. (1991). *Final Report: A descriptive study of significant features of exemplary special alternative instructional programs*. Los Alamitos, CA: Southwest Regional Educational Laboratory.

Table 1

Internal Reliability of the Six Empirically Derived Subscales of the Observation Instrument

	Subscale
Item Number	Item Description
Subscale 1: Explicit Teaching (6 items, alpha =.87)	
1	Models skills and strategies
2	Makes relationships overt
3	Emphasizes distinctive features of new concepts
4	Provides prompts
15	Length of literacy activities is appropriate
17	Adjusts own use of English during lesson
Subscale 2: Instruction Geared Towards Low Performers (7 items, alpha = .91)	
6	Achieves high level of response accuracy
7	Ensures quality of independent practice
10	Engages in ongoing monitoring of student understanding and performance
11	Elicits response from all students
12	Modifies instruction for students as needed
13	Provides extra instruction, practice and review
28	Asks questions to ensure comprehension
Subscale 3: Sheltered English Teaching Techniques (4 items, alpha = .78)	
18	Uses visuals and manipulatives to teach content
22	Provides explicit instruction in English language conventions
24	Encourages students to give elaborate responses

26	Uses gestures and facial expressions in teaching vocabulary and clarifying meaning of content
Subscale 4: Interactive Teaching (4 items, alpha = .65)	
8	Secures and maintains student attention during lesson
14	Extent to which students are "on task" during literacy activities
21	Selects and incorporates students responses, ideas, examples and experiences into lesson
23	Gives student wait time to respond to questions
Subscale 5: Vocabulary Development (4 items, alpha = .86)	
5	Teaches difficult vocabulary prior to and during lesson
20	Structures opportunities to speak English
27d	Provides systematic instruction of vocabulary development
29	Engages students in meaningful interactions about text
Subscale 6: Phonemic Awareness and Decoding (3 items, alpha = .87)	
27a	Provides systematic instruction in phonemic awareness[1]
27b	Provides systematic instruction in letter-sound correspondence
27c	Provides systematic instruction in decoding
Other Items	
9	Gives feedback on academic performance
16	Transitions between instructional activities are short and efficient
25	Use native language to help students understand content

[1] This item was dropped during the course of Year 2 because many of the teachers had completed instruction in phonemic awareness by the time the observations were conducted.

Table 2

Correlations Between Subscales: Construct Validity

			Sheltered			
	Explicit	Instruction for	English	Interactive	Vocabulary	Phonics and
Subscale	Teaching	Low Performers	Instruction	Teaching	Development	Decoding
1. Explicit Teaching	--					
2. Instruction for Low Performers	.825**	--				
3. Sheltered English Instruction	.598**	.493	--			
4. Interactive Teaching	.780**	.840**	.736**	--		
5. Vocabulary Development	.718**	.665**	.863**	.898**	--	
6. Phonics and Decoding	.769**	.600*	.223	.572*	.402	--

Note. Correlations are based on one rating per teacher (n = 14). * p < .05; ** p < .01

Table 3

Pretest and Posttest Performance of Students in Grade 1 by Language Group

			Native
	Spanish	Other Primary	English
	Speakers	Languages	Speakers
	n = 110	n = 31	n = 53
Student Measures	M (SD)	M (SD)	M (SD)
Pretests			
Letter Naming Fluency	40.43 (15.48)	49.55 (13.94)	48.00 (14.68)
Phonemic Segmentation Fluency	32.21 (19.26)	20.32 (14.67)	36.04 (17.39)
Nonsense Word Fluency	32.15 (16.01)	38.06 (28.20)	35.02 (17.92)
Posttests			
Nonsense Word Fluency	68.24 (33.72)	67.68 (29.29)	67.98 (28.01)
Oral Reading Fluency	50.37 (26.92)	73.97 (28.28)[a]	62.45 (30.02)
Reading Comprehension	3.91 (2.86)[b]	6.01 (2.98)[a]	5.31 (2.99)[c]

[a]n = 30; [b]n= 80; [c]n = 51.

Table 4

Correlations between Subscales on Observation Instrument and Composite Reading Score (with Class as Unit of Analysis)

Subscale	Correlation
1. Explicit Teaching / Art of Teaching	.75**
2. Instruction Geared Toward Low Performers	.60*
3. Sheltered English Techniques	.67**
4. Interactive Teaching	.62*
5. Vocabulary Development	.64*
6. Phonics and Decoding	.49
Total Observation score	.77**

* p < .05; ** p < .01..

Note. N = 14 classrooms.

Appendix A

English Language Learner Classroom Observation Instrument[1]

Instructional Practices

Items	***Field Notes***
1. Models skills and strategies during lesson 1 Not Effective · 2 Partially Effective · 3 Moderately Effective · 4 Very Effective	
2. Makes relationships among concepts overt 1 Not Effective · 2 Partially Effective · 3 Moderately Effective · 4 Very Effective	
3. Emphasizes distinctive features of new concepts • Broad range of examples and non-examples • Examples used to show relevant and irrelevant features 1 Not Effective · 2 Partially Effective · 3 Moderately Effective · 4 Very Effective	
4. Provides prompts and cues in how to use strategies, skills, and concepts • (e.g., guided practice, scaffolds, steps and procedures) 1 Not Effective · 2 Partially Effective · 3 Moderately Effective · 4 Very Effective	

[1] The first page of the instrument is represented in its final form. Subsequent items are presented without the accompanying rating scale.

5. Teaches difficult vocabulary prior to lesson, or during lesson as needed.

6. Achieves high level of response accuracy in context of lesson objectives.

    - (e.g., spelling accuracy on a spelling test vs. spelling accuracy on a written assignment.

7. Rate the quality of independent practice.

8. Secures and maintains student attention during lesson, as needed.

9. Gives feedback on academic performance.

- Reiterates, clarifies, reinforces.
- Communicates clearly what students did correctly or how they can improve.
- Focuses on lesson objective.

10. Focuses on performance specifics (i.e., not just "Good" or "Wrong").

11. Engages in *ongoing* monitoring of student understanding and performance *during* lesson.

- Elicits responses from all students, including students having difficulty with task at hand.
- Calls on range of students.
- Poses questions that students can answer.

12. Modifies instruction for students as needed during the lesson.

- Breaks down task into smaller/simpler components.
- Modifies assignments to promote success.
- Provides specialized instruction.

13. Provides extra instruction, practice, or review for students having difficulty with task at hand.

General Instructional Environment

14. Rate the extent to which students are "on-task" during literacy activities.

15. Length of literacy activities appears to be the right length for most students.

16. Transitions between instructional activities are short and efficient.

English-Language Development

17. Adjusts own use of English to make concepts comprehensible.

18. Uses visuals or manipulatives to teach content.

19. Gives oral directions that are clear and appropriate for level of students’ English language development.

20. Structures opportunities for students to speak.

21. Selects and incorporates students' responses, ideas, examples, and experiences into the lesson.

22. Provides explicit instruction in English language use, and includes the use of cue and prompts.

23. Gives students wait time to respond to questions.

24. Encourages students to give elaborate responses.

- Prompts students to expand on one-word or short answers.
- prompts student to provide more information.
- Prompts student to give more complete responses

25. The teacher and/or students strategically use students' native language to help students understand content.

26. Uses gestures and facial expressions in teaching vocabulary and clarifying meaning of content.

Content Specific to Reading/Language Arts

27. Provides systematic, explicit instruction in the following areas (each of these received it’s own rating):

- Phonemic awareness.
- Letter-sound correspondence.
- Decoding.
- Vocabulary and vocabulary development.

28. Checks student comprehension of text by asking questions.

29. Engages students in meaningful interactions about text.

Appendix B

Reading Comprehension Assessment Story

A Nice Bite of Rice

This is the home of a little brown dog. Dog's pals live down the lane. Her pals are a frog, a cat, and three white mice. "I need a new home," says Dog. "I will ask my pals to help." Dog sees Cat with a nice ball. "Will you help me make a home?" she asks. "No time!" says Cat. Dog sees the mice with a cone. "Will you help me make a home?" she asks. "No time!" say the mice. Dog sees Frog on a log. "Will you help me make a home?" she asks. "No time!" says Frog. "I could use help. You have no time to help a pal? You are not nice," says Dog. "Now I will have to make a home by myself." In time the home is made. It is quite a nice home. "Now I will dine," says Dog. "I will ask my pals to help me make rice.   Will you help me make rice?" asks Dog. "No time!" say Frog and Cat twice. "No time!" say the mice. "I could use help. Now I will have to make the rice by myself," says Dog. "You are not nice. I made nice rice!" says Dog. "But it is not much fun to be by myself. I will go and ask Frog, Cat, and the mice to come and dine on rice," says Dog with a smile. Frog, Cat, and the mice are not at home. Dog is very sad. She sobs and sobs. "I will go home and dine by myself," Dog sobs. Frog, Cat, and the mice are in Dog's home. "Surprise!" they yell. "Now we have time to help you. We came with bones and ice cream cones. We made the ice cream," say the mice. "We can dine with you and then we will do the dishes," says Cat. "This is quite nice!" says Dog with a smile. "I like to dine with my pals!"

Appendix B (continued)

Reading Comprehension Questions

1. What help did Dog want from her pals?
2. Why wouldn’t her pals help her?
3. Why does Dog want her pals to eat with her?
4. Why is Dog sad when her pals are not home?
5. What happens at the end of the story?

Appendix C

Deleted Items

- Reviews concepts, skills, and strategies.
- Gives feedback on academic performance: Quantity (retained quality for this item).
- Academic engagement is spent on speaking, writing, and reading.
- The classroom has a ‘tight’ academic feel.
- Students write in English.
- Models appropriate English language use.
- Gives oral directions that are clear and appropriate for level of students’ English language development.
- Conveys clear expectations about level and complexity of English language use expected.
- Reviews content of English language lessons.
- Rate the extent teacher-led instruction is devoted to: Whole class, small group (7 – 12 students), small group (2 – 6 students), one-to-one instruction.
- Rate the extent and quality of academic peer interactions.
- Written language is provided at appropriate level for students.

- Academic English is taught.
- Explicit instruction items in Reading/Language Arts section:
    - high-frequency sight words
    - word families/onsets-rimes
    - text structure/story mapping
    - spelling
    - mechanics of writing
    - concepts about print
- Reads aloud to students.
- Teaches requisite skills needed for reading and listening comprehension.
- Models good writing as part of instruction.
- Promotes student writing of phonetically plausible words, phrases, and sentences, and high frequency words.
- Provides opportunities for students to generate their own text (i.e., as opposed to copying).
- Uses decodable text during instruction.
- Checks students’ oral reading fluency.
- Provides instruction in the organizational structure of text.
- Engages in ongoing monitoring of student understanding during lesson.