Throughout this report the Fountas & Pinnell Sistema de evaluación de la lectura, Grados K–2, Niveles A–N, will be referred to as *Sistema*. It is the Spanish equivalent of the Fountas & Pinnell Benchmark Assessment System published by Heinemann in 2007 and revised in 2010.
**Assessment Summary Form**

List the titles read by the student from lowest to highest level.

<table>
<thead>
<tr>
<th>Title</th>
<th>Fiction/Nonfiction</th>
<th>Level</th>
<th>Accuracy</th>
<th>Comprehension</th>
<th>Independent*</th>
<th>Instructional**</th>
<th>Hard***</th>
<th>Self-Correction</th>
<th>Fluency</th>
<th>Rate</th>
<th>Writing About Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Independent Level**

Levels A–K: Highest level read with 95–100% accuracy and excellent or satisfactory comprehension.

Levels L–N: Highest level read with 98–100% accuracy and excellent or satisfactory comprehension.

**Instructional Level**

Levels A–K: Highest level read with 90–94% accuracy and excellent or satisfactory comprehension or 95–100% accuracy and limited comprehension.

Levels L–N: Highest level read with 95–97% accuracy and excellent or satisfactory comprehension or 98–100% accuracy and limited comprehension.

***Hard Level**

Levels A–K: Highest level read at which accuracy is below 90% with any level of comprehension.

Levels L–N: Highest level read at which accuracy is below 95% with any level of comprehension.

**Comprehension**

<table>
<thead>
<tr>
<th>Levels A–K</th>
<th>Levels L–N</th>
</tr>
</thead>
<tbody>
<tr>
<td>6–7</td>
<td>9–10</td>
</tr>
<tr>
<td>Excellent</td>
<td>Excellent</td>
</tr>
<tr>
<td>5</td>
<td>7–8</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>Satisfactory</td>
</tr>
<tr>
<td>4</td>
<td>5–6</td>
</tr>
<tr>
<td>Limited</td>
<td>Limited</td>
</tr>
<tr>
<td>0–3</td>
<td>0–4</td>
</tr>
<tr>
<td>Unsatisfactory</td>
<td>Unsatisfactory</td>
</tr>
</tbody>
</table>

Behaviors and Understandings to Notice, Teach, and Support (See [Continuo de adquisición de la lectoescritura](#))
INTRODUCTION

The Sistema de Evaluación de la Lectura (Sistema), Grados K–2, Niveles A–N is a formative assessment of reading in Spanish comprised of 28 high-quality original titles, or books, divided evenly between fiction and nonfiction. The Sistema measures decoding, fluency, vocabulary, and comprehension skills in kindergarten through third grade (mid-third grade). The set of books, recording forms, and other materials is an assessment tool for teachers, literacy specialists, and clinicians to use in determining students’ developmental Spanish reading levels for the purpose of informing instruction and documenting reading progress. The Sistema is the Spanish counterpart to the Fountas & Pinnell Benchmark Assessment System in English, published in 2007 (revised in 2010) to critical acclaim.

Since the Sistema is based on the Fountas & Pinnell Benchmark Assessment System, the results of the Spanish reading texts from this evaluation will be compared, where appropriate, with the results of the English reading texts collected in 2007. It should be noted that the Spanish-speaking student sample selected for the Sistema evaluation is a completely separate sample from the English-speaking sample selected for the Benchmark Assessment System.

In the field study with English readers in 2007, the Benchmark Assessment System was demonstrated to be a reliable measure of increased reading difficulty that is consistent with other Fountas and Pinnell leveled texts. If the Sistema is also a reliable measure of increased readability, it should produce similar reliability outcomes as those of the Benchmark Assessment System.

To determine if the Sistema is a valid and reliable assessment of a student’s Spanish reading level, a formative evaluation was conducted with a broad spectrum of classroom readers in different regions across the United States. This formative evaluation generated ongoing and immediate feedback from field-test examiners and Spanish readers in grades kindergarten through third grade that was used during the continued development of the program to ensure that it met standards of validity and reliability.

In summary, after two and one-half years of editorial development, field testing, and independent data analysis, the Sistema texts were demonstrated to be both reliable and valid measures for assessing students’ Spanish reading levels.

DESCRIPTION OF THE SISTEMA

The Sistema is aligned with the A–N book levels of the F&P Text Level Gradient™, and assesses basic reading competencies that are developed between kindergarten and early third grade. The Sistema books and their corresponding grade levels are depicted in Figure 1.

<table>
<thead>
<tr>
<th>F&amp;P TEXT LEVEL GRADIENT™</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fountas &amp; Pinnell LEVELS</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>H</td>
</tr>
</tbody>
</table>

figure 1
PURPOSE

A formative evaluation of the Sistema was conducted to ensure that (1) the leveling of the texts is reliable, (2) the reading scores are valid and accurately identify each student’s Spanish reading levels, and (3) fiction and nonfiction texts at each level are of comparable difficulty. The purpose of the evaluation study therefore was to measure the reliability and validity of Sistema.

To measure the reliability of the text levels, every book was examined within a broader developmental literacy framework and across corresponding fiction and nonfiction genres. The goal was to determine the readability of the books, and whether the readability was consistent across the fiction and nonfiction domains. For example, are the level G fiction and nonfiction books typical level G books, and are the corresponding fiction and nonfiction level G books at comparable levels of difficulty?

Finally, to measure the validity of the assessment outcomes, the Sistema was compared with other Spanish reading assessments. The purpose was to measure the extent to which the Sistema is associated with other valid Spanish reading assessments.

RESEARCH QUESTIONS

To determine the validity and reliability of the Sistema, the following three research questions guided the formative evaluation:

RESEARCH QUESTION 1: VERTICAL RELIABILITY

- How reliable is the Sistema? That is, how consistent and stable is the information derived from the Spanish books?

- Does each book of the Sistema consistently occupy the same position on the gradient of readability, based on multiple readings by age-appropriate students?

RESEARCH QUESTION 2: HORIZONTAL RELIABILITY

- To what extent are the gradients of difficulty for fiction and nonfiction books aligned within the Sistema?

- Do corresponding fiction and nonfiction books represent similar levels of difficulty within similar levels of Spanish reading?

RESEARCH QUESTION 3: CONVERGENT VALIDITY

- What is the test-retest reliability of the Sistema?

- What is the convergent validity between the Sistema and:
  - Evaluación del desarrollo de la lectura 2—EDL2 (Spanish version of the Developmental Reading Assessment, DRA2)
  - Text reading section of the Instrumento de observación de los logros de la lecto-escritura (Observation Survey in Spanish)
  - Lectura section of the Language Assessment Scales—LAS Links Español

METHODS

FORMATIVE EVALUATION

In order to determine validity and reliability of the Sistema, a project manager designed a formative evaluation of the program comparable to that used for the Fountas & Pinnell Benchmark Assessment System. Formative evaluation is a method of analyzing the effectiveness of a program in its development stages. For this evaluation, the field data were collected systematically and analyzed on an ongoing basis to ascertain the program’s attainment of its objectives. Interim reports were developed and utilized as a basis for determining the soundness, complexities, and utility of the program. Because the process incorporated ongoing feedback gathered by field-test examiners, the program authors and developers were able to make informed decisions regarding adjustments and
refinements. At the conclusion of the field study, an independent data analyst was brought in to evaluate the program’s validity and reliability.

This formative research was conducted in two phases. Phase 1 of the study addressed Research Question 1 (Does each book consistently and reliably represent a degree of increased difficulty in Spanish reading?) and Research Question 2 (Do corresponding Spanish fiction and nonfiction books represent similar levels of readability?). Phase 2 addressed Research Question 3 (To what extent is the Sistema associated with other established Spanish reading assessments?). Prior to the formative evaluation, an editorial process was used to establish the text leveling. This editorial developmental process is discussed next.

EDITORIAL DEVELOPMENT

BOOK DEVELOPMENT

Development of the texts for the Sistema was closely supervised by Dr. Irene Fountas and Dr. Gay Su Pinnell, creators of the F&P Text Level Gradient™, to ensure book development met their strict leveling protocols. Fountas and Pinnell worked as co-creators with a team of bilingual educators and researchers to design original texts that would reflect the language and culture of young Spanish readers. Team members were chosen for their experience in teaching with Fountas and Pinnell leveled books, as well as their experience teaching Spanish reading. Attention was paid to ensure the texts united relevant characteristics of the designated levels outlined in Leveled Books K–8: Matching Texts to Readers for Effective Teaching (Fountas & Pinnell, 2006) with attributes of well-written texts in Spanish. At every level, the Sistema books are distinguished by their culturally relevant topics, humor, language and illustrations. Each of the level A–N books, written and illustrated by Latino authors and artists, is 16 pages in length, and the challenges presented at each level increase incrementally. The Sistema provides the teacher, evaluator, or clinician an appropriate measure to assess a student’s Spanish reading accuracy, fluency, and comprehension level for informing instruction. While the Sistema provides supplemental measures of phonics, word analysis, high-frequency word assessment, vocabulary-in-context, and writing about reading, only the last of these subtests was trialed during the field study.

LEVELING BOOKS

A gradient of text is defined by Fountas and Pinnell (2006a) as “a varied collection organized into approximate levels of difficulty. Texts that increase demands in terms of concept, theme, vocabulary, length, and so on, are more difficult” (p. 84). The leveling team conferred on numerous occasions to determine the initial text levels. The program’s authors reviewed this initial leveling and made revisions to texts to arrive at a complete text set for field testing.

RESEARCH METHODS

RESEARCH PARTICIPANTS
STUDENTS

Students tested were included in the analyses provided they had opportunities to read texts in sequence (i.e., the examiners did not skip levels). Of the 166 students tested in Phase 1, 145 met this criterion; of the 164 students tested in Phase 2, 139 met the criterion. Nine (9) students read independent and instructional levels on fiction, but did not reach the hard level; twenty (20) students read independent and instructional levels on nonfiction, but did not reach the hard level. These 29 students were included in the final analyses, even though they lacked the hard level confirming that their instructional levels were indeed the highest levels they could read before reaching frustration. The final analyses were run with and without this cohort of 29 students that did not reach the hard level in fiction or nonfiction. Since the results with and without this cohort were nearly identical, and the reliability outcomes were not affected, these 29 students were included in the final results. The decision was made to include the 29 student cohort, because the affected students most likely would have reached hard at the next level had the testing not been stopped.
School sites from which these students were drawn were ethnically, socioeconomically, and geographically diverse. The research goal was to identify fluent Spanish-reading students making average progress in Spanish literacy instruction. Accordingly, students were selected on the basis of their ability to understand and speak Spanish, as well as rate of progress in their Spanish literacy. Field-test examiners tested for fluency in oral Spanish and conferred with classroom teachers to determine the student’s rate of literacy progress in the classroom.

FIELD-TEST EXAMINERS
Ten field-test examiners fluent and literate in Spanish were selected. All test field examiners were literacy educators who had extensive training in teaching young children to read and write in Spanish, in administering running records (Clay, 2002), and in using other forms of benchmark assessments to assess students’ reading levels. Seven of the ten examiners were either teachers or teacher leaders in Descubriendo la Lectura, the reconstruction of Reading Recovery™ in Spanish. Prior to the beginning of the field testing, a three-day intensive training session, led by project managers Jane Barboza and Raquel Mireles, was convened in August 2010 to train the field-test examiners in the formative evaluation’s protocols and procedures. The research specialist who had co-authored the field study report for the Fountas & Pinnell Benchmark Assessment System attended and participated in the training.

CONTEXTS
A total of seven schools participated in the Phase 1 field testing. The schools were located in five different geographic regions of the United States. Two field-test examiners were assigned to each of the five regions and conducted the field testing. Field testing took place in the following five regions:

Los Angeles County, CA
Chicago, IL
New York, NY
Denver, CO
San Antonio, TX

Phase 2 was conducted in the same five geographic regions, and at the same seven schools.

Because of the increasing diversity of student populations in today’s schools, schools that represented diverse socioeconomic settings (SES) were targeted. These determinations were made by referring to Federal guidelines for student participation in the free and reduced lunch program, used by school districts as an indicator of poverty (see Figure 2).

To evaluate the reliability and validity of the Sistema, it was important that the schools have a large and representative enrollment of students learning to read in Spanish. Accordingly, all participating schools served an English language learner enrollment of at least 36% (see Figure 2).

<table>
<thead>
<tr>
<th>SCHOOL FIELD SITES</th>
<th>California</th>
<th>New York</th>
<th>Texas</th>
<th>Illinois</th>
<th>Colorado</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of school sites (elementary and middle school)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Average percentage of students receiving free or reduced-price lunch or economically disadvantaged</td>
<td>73.9%</td>
<td>93.0%</td>
<td>84.0%</td>
<td>98.5%</td>
<td>87.0%</td>
</tr>
<tr>
<td>Average percentage of English Language Learners</td>
<td>45.1%</td>
<td>38.0%</td>
<td>36.0%</td>
<td>42.2%</td>
<td>50.7%</td>
</tr>
</tbody>
</table>
In summary, students in Phases 1 and 2 represented a cross-section of the major regions of the U.S., diverse socioeconomic levels and language backgrounds, and participated in programs for Spanish literacy instruction. Figures 2 and 3 depict this information.

Figure 3 shows the average percentage of students by ethnicity from the field site schools from each state.

<table>
<thead>
<tr>
<th>ETHNICITY OF STUDENTS BY SCHOOL FIELD SITES</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>African American</td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td>Asian American &amp; Pacific Islander</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
</tr>
<tr>
<td>Multiracial/Other</td>
</tr>
</tbody>
</table>

**PROCEDURES AND MATERIALS**

**PHASE 1 OF THE FORMATIVE EVALUATION**

Phase 1 of the study examined Research Questions 1 and 2, which respectively addressed the consistency of the vertical gradient of each level, and the horizontal gradient within each level for both fiction and nonfiction books. The Sistema books were tested in the following sequence:

1. Fiction
2. Nonfiction

**PROCEDURES FOR ASSESSMENT ADMINISTRATION**

Reading data for every student using both fiction and nonfiction books were gathered systematically through a formative evaluation design protocol. After the August 2010 training, the field-test examiners began working individually at the selected school sites during the last quarter of 2010. By testing for oral Spanish fluency and conferring with classroom teachers at each school site, field-test examiners identified eligible students who met the criteria for inclusion in the study (i.e., students who were fluent Spanish speakers and average-progress literacy learners). Below are descriptions of protocols and procedures followed by each field test examiner.

**DETERMINING FLUENCY IN SPEAKING AND UNDERSTANDING SPANISH**

Examiners administered the grade level-appropriate version of the Hablando subtest of the *LAS Links Español* to every prospective subject—that is, average-progress kindergarten, first, second and third grade students. (Third grade students were included in both phases of the testing in order to obtain fluent readings of upper-level texts.) Students scoring in the Proficient and Above Proficient ranges were deemed eligible for selection as subjects. Once scores were tabulated, examiners confirmed their findings with classroom teachers, selected 12–18 students, and obtained parental consent for inclusion in the field study.

**SELECTING A STARTING POINT FOR READING**

An “Antes de empezar” (Where to Start) word list was developed by the program’s authors to assist examiners in quickly placing students at their appropriate Spanish reading levels. This word list was administered to each subject at the beginning of individual testing. Using this as a starting point, and taking into account data from classroom teachers, the field-test examiners determined which book they should ask each student to read first.
DETERMINING INDEPENDENT, INSTRUCTIONAL, AND HARD SPANISH READING LEVELS

Next, field-test examiners assessed each student’s ability to read and comprehend at least three sequential levels of books in the fiction genre. Specifically, the field-test examiners sought to identify one book for each student that was relatively easy (i.e., the student’s independent reading level); one book that offered just enough difficult vocabulary and/or concepts to make the reading interesting and challenging (i.e., the student’s instructional reading level); and a third book that was too challenging to be rewarding (i.e., the student’s hard reading level).

After each book was read aloud, student and examiner engaged in a comprehension conversation for that particular book. Examiners referred to prompts printed on the Recording Forms, but they were encouraged to paraphrase or re-phrase prompts as needed to move the conversation along or probe more deeply for understanding. Examiners rated students’ understanding of a text using the Fountas and Pinnell comprehension guidelines (2011, p. 25-30). The focal areas listed below were rated on a scale from 0–3:

- Thinking within the text;
- Thinking beyond the text; and
- Thinking about the text.

Students who demonstrated one or more unique or valuable understandings were awarded an extra point.

Independent, instructional and hard levels were determined by taking into account both accuracy and comprehension. In every case, examiners consulted the following matrix depicted in Figure 4.
In some instances, a student’s scores did not exactly match the criteria for one or more of the three levels. In these cases, the examiner, in conference with the project manager, used additional information about the student in order to make a decision. To make these determinations, reference was made to Assessment Guide: A Guide to Benchmark Assessment System 1, pages 45-48 (2011).

Writing About Reading

Writing About Reading is an optional assessment within both the Benchmark Assessment System and Sistema. During Phase 1, examiners administered this subtest to students on the book determined to be at their instructional level to gather additional information about their understandings. In the context of classroom instruction, writing yields valuable information about a student’s processing of the printed language, as well as understanding and engagement with text.

Assessing Fluency

As Pinnell, Pikulski, Wixson, Campbell, Gough, and Beatty (1995) point out, fluency is an indicator of students’ understanding of text. It is expected that students should read along at a reasonable pace when reading at their instructional level. Examiners rated students’ fluency using a four-point rubric (Assessment Guide: A Guide to Benchmark Assessment System 1, p. 4 2011). The fluency score does not figure in calculating independent, instructional and hard levels.

Determining the Corresponding Readability Between Fiction and Nonfiction Books

Field-test examiners followed the process described above, using both fiction and nonfiction books. Given that students’ reading levels had been established with fiction texts, the field-test examiners did not need to re-administer the Antes de empezar word list test for nonfiction.

Anticipating Varying Developmental Reading Patterns

The research project manager and program developers were aware that ascertaining students’ three sequential reading levels could be a more complex process than the one outlined above. They fully anticipated varying developmental levels and an up-and-down pattern in a child’s reading of progressively more difficult texts. These possibilities were covered extensively during the training session for the field-test examiners. Such patterns could be attributed to a variety of factors, such as classroom instructional emphasis, students’ interest in subject matter, motivation, need for warm-up time, or reader fatigue, among other explanatory factors, all of which are beyond the scope of this study. To support ongoing data results, the research project manager provided additional support relative to procedures and particular cases throughout the testing process.

Schedule of Assessment Administration

The field-test examiners worked on a staggered schedule between September 13 and December 4, 2010. This allowed them to refine the day-to-day practical aspects of the research as needed and to suggest modifications to books that posed unforeseen challenges to young readers.

In general, the schedule flowed as indicated in Figure 5. This schedule was altered for the five examiners who were full-time classroom teachers and extended for all to allow more time to collect and review data.
### FIVE-DAY SCHEDULE FOR FIELD TEST EXAMINERS, PHASE 1

**DAY 1:**
- Administer LAS Links español—Hablando
- Administer to average-progress kindergarten through third grade students

**DAY 2:**
- Complete and score LAS Links español—Hablando
  - In conference with classroom teachers, select 18 students
- Obtain parent consent
- Prepare for testing

**DAY 3:**
- Administer Sistema fiction to 12 students
  - For each student, find independent, instructional and hard levels on fiction texts

**DAY 4:**
- Administer Sistema nonfiction to same 12 students
  - For each student, find independent, instructional and hard levels on nonfiction texts

**DAY 5:**
- Administer Sistema fiction and nonfiction to last 6 students

---

### FIELD TESTING DOCUMENTATION

Given the complexity of the assessment process, field-test examiners were responsible for maintaining ongoing detailed records of their findings related to the student’s readings and writing. Documentation was completed on the Sistema Assessment Summary form (see inside cover for sample). For each student, examiners compiled a list of fiction and nonfiction texts read, setting out independent, instructional and hard levels, and noting accuracy, comprehension, fluency, writing scores, and any particular concerns. These lists, as well as every reading record and writing sample accompanying them, were reviewed by the project manager for accuracy.

### RESEARCH DEBRIEFINGS

On an as-needed basis, the field-test examiners analyzed new data and reported back to the research project manager in debriefings by phone and email. These debriefings provided an opportunity to take action on revising texts to address unwanted anomalies in reading and understanding texts.

---

### FORMATIVE SYSTEM DEVELOPMENT

With a formative evaluation process, data analysis was ongoing as well as recursive. Ongoing field data gathered in authentic contexts provided information for adjustments and revisions in the system. Based on these data, the research project manager identified patterns both within individual books and across books using the following three categories of text evaluation:

1. Texts that were completely on-target;
2. Texts that required minimal changes or revisions;
3. Texts that required substantive revisions

None of the books had to be replaced altogether.

In December 2010 and January 2011, based on the Sistema testing results, changes were made in some of the texts. For example, students in New York City had difficulty recognizing an illustration of a one-story school. Creators of the book altered the illustrations accordingly. Other changes included shortening of some texts, altering introductions, or tightening the correspondence between text and illustration. The revised texts were trialed with students during Phase 2 of the field study.
When Phase 1 of the field testing was completed, the research specialist, experienced in quantitative data analysis as well as research design, methods, and data collection, analyzed the data. This was the same research specialist who helped evaluate the Fountas & Pinnell Benchmark Assessment System © 2008, 2011. Since the Sistema is based on the Fountas & Pinnell Benchmark Assessment System, the results of the Spanish reading texts from this evaluation will be compared, where appropriate, with the results of the English reading texts collected in 2007. The Benchmark Assessment System was demonstrated to be a reliable measure of increased reading difficulty that is consistent with other Fountas and Pinnell leveled texts. If the Sistema is also a reliable measure of increased readability, it should produce similar reliability outcomes as those of the Benchmark Assessment System. Once again, it should be noted that the Spanish-speaking student sample selected for the Sistema evaluation is a completely different sample from the English-speaking sample selected for the Benchmark Assessment System.

PHASE 1 OF THE FORMATIVE EVALUATION

Phase 1 of the study examined Research Questions 1 and 2, which related to the vertical gradient level for both fiction and nonfiction books, as well as the corresponding consistency of readability between fiction and nonfiction books. The results of Phase 1 are divided into two sections. The first section addresses Research Question 1 and the second section addresses Research Question 2.

SECTION 1. RELIABILITY OF VERTICAL TEXT GRADIENT

RESEARCH QUESTION 1: VERTICAL RELIABILITY

- How reliable is the Sistema? That is, how consistent and stable is the information derived from the Spanish leveled books?

- Does each book of the Sistema consistently occupy the same position on the gradient of readability, based on multiple readings by age-appropriate students?

The findings, obtained from field testing conducted in varied geographic regions throughout the country, demonstrate that the Sistema books do indeed become incrementally more difficult as the levels progress vertically from Levels A–N.

SECTION 1. DATA ANALYSIS OF VERTICAL TEXT GRADIENT

All students with complete data were included in the analysis. Students with complete data were those who had an instructional level and that had also been tested on the books immediately preceding and succeeding the instructional level were included. Students were not included if they had not tested at an instructional level, or if the books were too hard (a reading level of A was not established) or the books were too easy (a reading level of N was read at the independent level). Of 166 total students tested, 145 met the criteria for inclusion in the study.

SECTION 1. FINDINGS OF VERTICAL TEXT GRADIENT

Students read leveled texts in sequential and hierarchical progression. That is, they read the texts from lower to higher levels of text difficulty. The students’ reading progression from lower levels on the F&P Text Level Gradient™ to higher levels was sequential and hierarchical. The independent level, instructional level, and hard level were in the expected order of the text gradient. For example, when level D was the instructional level, then C was less difficult than D, and E was more difficult than D, as illustrated in Figure 6.
SECTION 1. RESULTS OF THE VERTICAL TEXT GRADIENT

For the Sistema, 90.3% of the students read the fiction books and 90.6% read the nonfiction texts in sequential and hierarchical order. Figure 7 displays the data confirming that texts became more difficult as a reader progressed through them in sequence. These results are similar with the outcomes collected with the Benchmark Assessment System in 2007.

As noted in Figure 7, most students generally read the Sistema fiction books and nonfiction books in sequential order.

Figures 8 and 9 depict the percentage of students who read in a sequential and hierarchical order from lower to higher levels of Spanish reading difficulty based on text level. That is, these are the proportion of cases in which students read the books as part of a consecutive progression from independent to hard.
At every level, the percentage of students who read the fiction books in sequential order was near 80% and above. In five of the fourteen levels (levels A, C, E, K, and N), all (100%) the students who met the testing criteria read fiction books in sequential order. At two other levels (levels B and M), 90% or more of the students read the fiction books in sequential order.

For the nonfiction books, the percentage of students who read the books in sequential order was 80% and above. In four of the fourteen levels (levels A, B, D, and N), all (100%) the students who met the testing criteria read nonfiction books in sequential order. At three other levels (levels C, E, and M), 90% or more of the students read the nonfiction books in sequential order.

**SISTEMA FICTION AND NONFICTION BY GRADE LEVEL**

Figure 10 represents the progress of students reading the Sistema fiction and nonfiction books (levels A–N) in sequential order, by grade level. The chart shows that at all grade levels (K–3), a similar percentage of students read the Sistema fiction books and nonfiction in sequential order at each grade level. Grade 3 results reflect reading level through mid-third grade. (Every grade-level percentage for Sistema is at least as high as the percentages established for the Benchmark Assessment System.)

**SECTION 2. RELIABILITY OF THE HORIZONTAL MATCH BETWEEN FICTION AND NONFICTION TEXTS**

In addition to answering the first research question in which the vertical text gradient was examined, Phase 1 also answered the study’s second question.

**PERCENTAGE OF STUDENTS READING FOUNTAS AND PINNELL LEVELED BOOKS IN SEQUENTIAL ORDER**

<table>
<thead>
<tr>
<th>Sistema Book</th>
<th>Kindergarten</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction</td>
<td>97.6%</td>
<td>88.8%</td>
<td>89.1%</td>
<td>92.8%</td>
</tr>
<tr>
<td>Nonfiction</td>
<td>97.7%</td>
<td>90.7%</td>
<td>87.3%</td>
<td>90.9%</td>
</tr>
</tbody>
</table>

**RESEARCH QUESTION 2: HORIZONTAL RELIABILITY**

- To what extent are the gradients of difficulty for fiction and nonfiction books aligned within the Sistema?
- Do corresponding fiction and nonfiction books represent similar levels of difficulty within the same levels of Spanish reading throughout the gradient of difficulty?

This section includes a horizontal analysis of fiction and nonfiction books at each level to determine if they are at the same degree of difficulty. For example, is a level D fiction book at the same level of difficulty as a level D nonfiction book?

The findings, obtained from field-testing conducted in the varied geographic regions throughout the country, indicate that the fiction and nonfiction books are indeed at similar levels of difficulty at each level of the A–N text gradient.

**DATA ANALYSIS OF HORIZONTAL TEXT CONSISTENCY**

All students with complete data were included in the analysis. Only the 284 students (145 students from Phase 1 and 139 from Phase 2) that had an instructional level in both fiction and nonfiction were included.

**RESULTS OF HORIZONTAL TEXT CONSISTENCY**

A preponderance of students read the texts gradient in one of two ways. Students either read fiction and nonfiction texts at the same level on the A–N text gradient, or they read fiction and nonfiction texts at similar levels of difficulty on the A–N text gradient. The following describes each.
I. FICTION AND NONFICTION TEXTS REPRESENT SAME LEVEL OF TEXT DIFFICULTY

The student’s developmental Spanish reading level is the same for fiction and nonfiction on the A–N text gradient. That is, the students’ instructional level in fiction is the same as in nonfiction. For example, a student’s instructional level is level D for both fiction and nonfiction. For Sistema, 47.7% of the students read at the same level (exact match) in fiction and nonfiction (Figure 11).

II. FICTION AND NONFICTION TEXTS REPRESENT SIMILAR LEVEL OF TEXT DIFFICULTY

The second way students read was at a similar level for fiction and nonfiction. Students’ instructional levels on the fiction and nonfiction texts varied by one level of difficulty on the A–N gradient. For example, a student reading on an instructional level D on a fiction text would read on an instructional level of C (preceding level) or E (succeeding level) on the nonfiction text. Figure 11 illustrates that 40.7% of the students read the Spanish fiction and nonfiction books at a similar level (text within one level).

For the Sistema, 88.4% of the students read at either the same or similar instructional levels on both fiction and nonfiction texts (Figure 11).

<table>
<thead>
<tr>
<th>Sistema Book</th>
<th>Reading at Same Level (Exact Match) Between Fiction and Nonfiction</th>
<th>Reading at Similar Level (Within 1 Level) Between Fiction and Nonfiction</th>
<th>Reading at Same or Similar Level Between Fiction and Nonfiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>47.7%</td>
<td>40.7%</td>
<td>88.4%</td>
</tr>
</tbody>
</table>

RELATIONSHIP BETWEEN SISTEMA AND BENCHMARK ASSESSMENT SYSTEM

The relationship between the Sistema and the Benchmark Assessment System needs clarification. Overall, the findings for both systems are very similar, even though the data for each series were collected three years apart and with different sets of subjects. As with the Fountas & Pinnell Benchmark Assessment System, the Sistema books become progressively more difficult as the levels progress vertically from levels A–N (vertical consistency or reliability), and the fiction and nonfiction texts are read at the same or similar levels of difficulty (horizontal consistency or reliability). Both systems are vertically and horizontally consistent (reliable).

The research literature has consistently demonstrated over the past 30 years that reading and literacy skills acquired in one language will transfer to another language (Carlo, 2009; Cummins, 2002; Salazar, 2008). That is, students who are likely to read at grade level in Spanish are also more likely to read at grade level in English after a transitional period from Spanish reading to English reading is completed. What this means for the Sistema is that students who are able to read the books at any particular instructional level in Spanish will eventually transfer these skills and read with good comprehension in English, provided they are given instruction that supports them in this transition. However, it is very important to understand that the Sistema does not identify instructional reading levels in English. Only the Benchmark Assessment System, designed and field-tested with students in English-medium instruction, is appropriate to measure a student’s instructional level in English.

PHASE 2 OF THE FORMATIVE EVALUATION

Phase 1 of the study examined Research Questions 1 and 2, which related to the vertical gradient level and horizontal reading consistency between fiction and nonfiction books. Phase 2 of the formative evaluation examined Research Question 3, addressing the test-retest reliability and convergent validity of the Sistema with other measures of Spanish reading.

RESEARCH QUESTION

Phase 2 answered the final research questions.

RESEARCH QUESTION 3: CONVERGENT VALIDITY

• What is the test-retest reliability of the Sistema?

• What is the convergent validity between the Sistema and:
  • Evaluación del desarrollo de la lectura 2, EDL2 (Spanish version of the Developmental Reading Assessment, DRA2)
  • Text reading section of the Instrumento de observación de los logros de la lecto-escritura (reconstruction in Spanish of the Observation Survey)
  • Lectura section of the Language Assessment Scales–LAS Links Español
RELIABILITY

Reliability addresses the consistency of scores of an assessment. Test-retest reliability refers to the consistency and stability of scores obtained by the same person when examined with the same test on different occasions or with different sets of equivalent test items. To measure the test-retest reliability of Sistema, the students’ reading scores on the fiction series were correlated with their scores on the nonfiction series. An assumption underlying this study is that students who attain a given level on the fiction texts will perform similarly when reading the nonfiction texts.

This test-retest reliability is similar to the analysis conducted in Phase 1 regarding the percentage of students who read the fiction and nonfiction Sistema books at the same or similar levels, but a percentage is not a test of reliability. A test-retest reliability correlation measures the consistency of an instrument or assessment from one time to another. In Phase 1, we measured the percentage of students reading at the same or similar levels. In Phase 2 we evaluated the consistency of the Sistema assessment as a reliable measure of Spanish reading. In Phase 1 the focus was on student performance; in Phase 2 the focus was on the reading assessment, independent of student performance.

According to Nunnally and Bernstein (1994), three benchmarks exist for measuring a reliability coefficient, and thereby determining whether an assessment is stable, consistent, and dependable. A reliability coefficient value of .70 is sufficient for the early stages of research such as a pilot study. Basic research, such as the Sistema field study, should require test scores to have a reliability coefficient of .80 or higher. However, when important decisions are to be made based on test scores, such as identifying a student’s reading level, Nunnally and Bernstein claim that a reliability coefficient of .90 is the minimum, with .95 or higher a desirable standard.

As the test-retest results depicted below (Figure 12) demonstrate, both the Sistema and Benchmark Assessment System meet the most stringent reliability criteria. Both are stable, consistent, and dependable assessments for determining a student’s reading level.

CONVERGENT VALIDITY

The validity of the test is the degree to which the assessment measures what it purports to measure. Validity is a check on how well an assessment fulfills its stated function. Convergent validity examines the relationship between test scores among assessments that measure similar variables. Therefore, the assessment outcomes from the Sistema texts should be related with similar tests that assess Spanish reading. To address the convergent validity, the following three questions were addressed:

- What is the convergent validity between the Sistema and:
  - EDL2–Evaluación del desarrollo de la lectura (Spanish version of the DRA2)?
  - Text reading section of the Instrumento de observación de los logros de la lecto-escritura (reconstruction in Spanish of the Observation Survey)?
  - Lectura subtest of the Language Assessment Scales (LAS Links Español)?

METHOD

In Phase 2, the ten field-test examiners once again followed the same procedures described in Phase 1 to determine students’ independent, instructional, and difficult levels of reading proficiency at grade levels (K–3) on the Sistema texts. Next, field-test examiners administered the text reading section of the Instrumento de observación de los logros de la lecto-escritura to students in grades K–3. Finally, five of ten field-test examiners administered the Evaluación del desarrollo de la lectura 2 (EDL2), while the other five
administered the *LAS Links Español* (Lectura). As in Phase 1, field-test examiners systematically maintained data records and participated in debriefings as needed with the research project manager.

**DESCRIPTION OF ASSESSMENTS**

The *Instruemento de observación de los logros de la lecto-escritura* (1996) is the reconstruction in Spanish of the *Observation Survey of Early Literacy Achievement* (1993). The *Instruemento* consists of six literacy tasks, one of which is the *Análisis actual del texto*, or Text Reading. This task “involves the reading of whole texts and the taking of running records related to text reading” (p. 31 Clay et. al., 1996). Increasingly difficult texts are used to ascertain each student’s independent, instructional and hard reading levels. In recent studies with English readers, the Text Reading Level of the *Observation Survey* was correlated with other standardized, norm-referenced tests. These include the *Iowa Test of Basic Skills* (.764 correlation) (Gómez-Bellengé, Rogers, Wang, & Schulz, 2005) and the *Gates-MacGinitie Reading Test*, the *Woodcock Reading Mastery Test* (Gómez-Bellengé & Thompson, 2005). A corresponding study comparing *Sistema* with other measures of Spanish reading was needed.

The *Evaluación del desarrollo de la lectura* (*EDL2*), by Olivia Ruiz and Vivian Cuesta, is the adaptation in Spanish of the *English Developmental Reading Assessment K–8, Second Edition*, by Joetta Beaver and Mark Carter. It is designed to be used in kindergarten through sixth grade bilingual or dual-language classrooms, to help “teachers determine each student’s independent reading level and identify what needs to be learned next” (p. 4). Like the text reading section of the *Instruemento de observación*, the *EDL2* provides texts of incrementally increasing difficulty. The child reads the text while the teacher takes a running record and notes various reading behaviors and responses. The teacher then uses a continuum provided for each text to determine the level of difficulty of the text—intervention/emerging, instructional/developing, independent, or advanced. For purposes of this study, scores were compiled based upon students’ oral reading fluency, established by measures of expression, phrasing, rate, and accuracy. For levels 4 and above, a score of 11-16 was considered independent; between 7 and 10, instructional; and 6 or below, hard. For Levels 3 and below, scores were compiled based upon students’ oral reading, consisting of monitoring/self-corrections, use of cues and accuracy. At these lower levels, where the maximum number of points a child can score is nine (9), criteria for independent, instructional, and hard were adjusted downward.

The *Lectura* section of the *LAS Links Español* is a group-administered test consisting of three sections—*Analizar palabras*, *Leer palabras*, and *Leer para entender*. In *Analizar palabras*, students are asked to select words that contain same sounds or rhyme with given target words. In *Leer palabras*, students find words in response to oral or written cues. *Leer para entender* provides passages followed by multiple choice comprehension questions. Depending on their grade levels, students were tested on either the Grado K–1 or Grado 2–3 versions, individually or by grade-level groups.

**RESULTS**

**CONVERGENT VALIDITY: SISTEMA AND INSTRUMENTO**

Figure 13 shows a strong relationship between the reading accuracy rates on *Sistema* fiction (correlation of .88) and nonfiction (correlation of .87), and reading accuracy rates on the *Instruemento* (Spanish reconstruction of the *Observation Survey*). In other words, the performance on the *Sistema* texts is strongly indicative of performance on the *Instruemento* text reading assessment. These results reinforce the validity of the *Sistema* as a measure of reading in Spanish.

[It is interesting to note that this finding is similar to the strong relationship between the *Observation Survey* and *Benchmark Assessment System* fiction (correlation of .94) and nonfiction books (correlation of .93) reported three years previously.]
CONVERGENT VALIDITY WITH EVALUACIÓN DEL DESARROLLO DE LA LECTURA 2 (EDL2)

Another aspect of the Phase 2 formative evaluation examined the relationship between the Sistema and the EDL2 — Evaluación del desarrollo de la lectura (Spanish version of the DRA2 — Developmental Reading Assessment). The results, shown in Figure 13, indicate that performance on the Sistema fiction (correlation of .91) and nonfiction (correlation of .89) texts is strongly indicative of performance on EDL2 Spanish reading. Since the EDL2 is a developmental reading assessment, this strongly indicates that the Sistema is also a valid measure of progressive reading skills.

CONVERGENT VALIDITY WITH LAS LINKS ESPAÑOL (LECTURA)

The final analysis of the Phase 2 formative evaluation gauged the relationship between the Sistema and the Lectura section of the LAS Links Español. As reflected in Figure 13, the Sistema fiction and nonfiction books were moderately related with performance on the Lectura section of the LAS Links Español (correlations of .45 and .50 respectively). These results therefore show that the Sistema is moderately indicative of performance on the Lectura subtest of LAS Links Español. This lower correlation is not surprising in light of the fact that the concepts tested on the Lectura subtest of the LAS Links Español differ substantially from those tested on Instrumento, EDL2 and Sistema. LAS Links focuses on isolated words and sounds whereas Benchmark Assessment System measures reading of continuous text.

RESEARCH SUMMARY—PHASES 1 AND 2

Research Question 1 asked if each Spanish book from the Sistema for levels A–N represented a degree of increased difficulty that is consistent with other Fountas and Pinnell leveled texts.

- Analysis of the field testing indicates that relative to the text gradient, the Sistema Spanish books become incrementally more difficult as the levels progress vertically from levels A–N.

- For the Sistema Spanish reading books in grades K–3, 90.3% of the students read the fiction texts and 90.6% read the nonfiction books in a sequential and hierarchical order.

Research Question 2 addressed the extent to which the gradients of difficulty for fiction and nonfiction books were aligned within the Sistema. That is, do fiction and nonfiction books represent similar levels of difficulty within similar levels of reading?

- For Sistema (grades K–3), 88.4% of the students read at same or similar levels in fiction and nonfiction.

Research Question 3 addressed the test-retest reliability and convergent validity of the Sistema with other assessment measures.

- Test-retest reliability for the Sistema was .96.

- There was a strong association between the Sistema fiction (correlation of .88) and nonfiction (correlation...
of .87), and Instrumento, the Spanish version of the Observation Survey. [This finding is similar to the strong relationship between the Benchmark Assessment System fiction (correlation of .94) and nonfiction books (correlation of .93), and Observation Survey reported three years previously.]

- The results indicate a strong correlation between performance on the Sistema fiction (correlation of .91) and nonfiction (correlation of .89) texts and EDL2 — Evaluación del desarrollo de la lectura, a widely-used Spanish reading assessment.

- The Sistema fiction and nonfiction books were moderately indicative of performance on Lectura section of the LAS Links Español (correlations of .45 and .50, respectively)

In conclusion, the Fountas & Pinnell Sistema de evaluación de la lectura, Grados K–2, Niveles A–N was demonstrated to be both a reliable and valid measure for assessing student’s Spanish reading levels, showing high convergent validity with two other measures of Spanish reading and moderate convergent validity with a third measure. With its original books by Spanish-speaking authors and illustrators, its calculation of comprehension in the determination of reading levels, ease of use, and accompanying materials, it will provide accurate and useful information to teachers of students learning to read in Spanish.
BIBLIOGRAPHY


Salazar, J. J. (2008). Reading Outcomes of English Language Learners: Long-Term Affects of First Language (L1) and Second Language (L2) Proficiency, and Early L1 and L2 Reading Skills. Saarbrucken, Germany: VDM Verlag Dr. Müller Aktiengesellschaft & Co. AG.