

Field Study of Reliability and Validity of the Fountas & Pinnell Benchmark Assessment Systems 1 and 2

INTRODUCTION

The *Fountas & Pinnell Benchmark Assessment System* is a formative reading assessment comprising 58 high-quality original titles, or “little books,” divided evenly between fiction and nonfiction. The assessment measures decoding, fluency, vocabulary, and comprehension skills for students in kindergarten through 8th 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 reading levels for the purpose of informing instruction and documenting reading progress.

To determine whether the *Fountas & Pinnell Benchmark Assessment System* is a valid assessment of a student’s 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 readers that was used during the continued development of the program to ensure that it met standards of reliability and validity.

In summary, after two and a half years of editorial development, field testing, and independent data analysis, the *Fountas & Pinnell Benchmark Assessment System* texts were demonstrated to be both reliable and valid measures for assessing students’ reading levels.

DESCRIPTION OF THE BENCHMARK ASSESSMENT SYSTEM

The *Fountas & Pinnell Benchmark Assessment System* is aligned with the A–Z book levels of the Fountas & Pinnell Leveled Text Gradient. *System 1* represents levels A–N on the Fountas & Pinnell Text Gradient and encompasses kindergarten through grade 2. *System 2* represents Levels L–Z on the Fountas & Pinnell Text Gradient and encompasses grades 3 through 8. Recognizing the critical junctures in a child’s literacy development between grade 2 and grade 3, the *Fountas & Pinnell Benchmark Assessment System* levels L, M, and N offer twice as many books (four books per level). The representations of books in *Benchmark Systems 1 & 2* and their corresponding grade levels are depicted in Figure 1.

PURPOSE

A formative evaluation of the *Fountas & Pinnell Benchmark Assessment System* was conducted to ensure that (1) the leveling of the texts is reliable and (2) the reading scores are valid and accurately

identify each student’s reading level.

The purpose of the study was twofold. The first was to examine every book, at every level, for the reliability of its designated level within a broader literacy framework and across corresponding fiction and nonfiction genres. That is, is the readability of the books consistent across the fiction and nonfiction domains? For example, are the level G fiction and nonfiction books not only *typical* level G books, but do corresponding fiction and nonfiction books at this level have the same degree of readability?

The second purpose of the evaluation was to determine the correlation between the *Fountas & Pinnell Benchmark Assessment System* and other reading assessments. That is, to what extent is the *Fountas & Pinnell Benchmark Assessment System* associated with other valid reading assessments?

RESEARCH QUESTIONS

In order to determine the reliability and validity of the *Fountas & Pinnell Benchmark Assessment System*, the following three research questions guided the formative evaluation:

Research Question 1

- How reliable is the *Fountas & Pinnell Benchmark Assessment System*? That is, how consistent and stable is the information derived from the reading books?
- Does each book of the *Fountas & Pinnell Benchmark Assessment System* consistently occupy the same position on the gradient of readability, based on multiple readings by age-appropriate students? That is, does each book, level A–Z represent

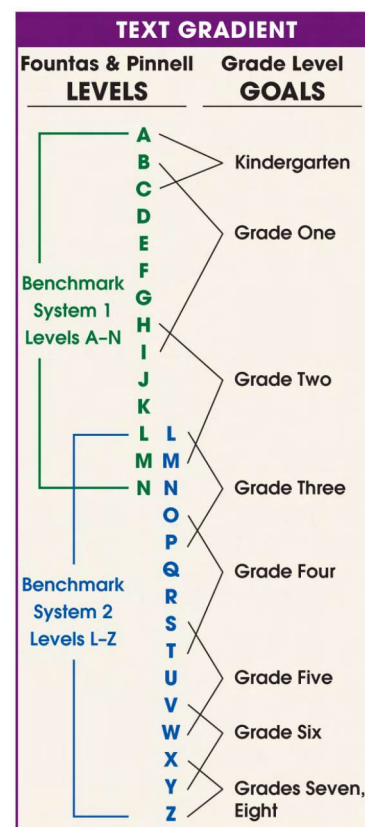


Figure 1

a degree of increased difficulty that is consistent with other Fountas and Pinnell leveled texts?

Research Question 2

- To what extent are the gradients of difficulty for *fiction* and *nonfiction* books aligned within the *Fountas & Pinnell Benchmark Assessment System*? Do fiction and nonfiction books represent similar levels of difficulty within similar levels of reading?

Research Question 3

- To what extent is the *Fountas & Pinnell Benchmark Assessment System* associated with other established reading assessments?
 - What is the convergent validity between the *System 1* and Reading Recovery[®] assessment texts?
 - What is the convergent validity between the *System 2* and the Slosson Oral Reading Test—Revised (SORT-R3) and the Degrees of Reading Power[®] (DRP)?

METHODS

Formative Evaluation

In order to determine reliability and validity, a research project manager designed a formative evaluation of the program. Formative evaluation is a method of analyzing the effectiveness of a program in its development stages. In this evaluation of the *Fountas & Pinnell Benchmark Assessment System*, 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 used 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-analysis team was brought in to evaluate the program's reliability and validity.

This formative research was conducted in two phases. Phase I of the study addressed research questions 1 and 2; Phase II addressed research question 3. Prior to the formative evaluation, an editorial process was used to establish the text leveling. This editorial development process is discussed next.

EDITORIAL PROGRAM DEVELOPMENT

Book Development

Development of the texts for the *Fountas & Pinnell Benchmark Assessment System* was closely supervised by Irene Fountas and Gay Su Pinnell, creators of the A–Z Text Gradient, to ensure book development met their strict leveling protocols. Attention was paid to ensure the texts reflected the specific characteristics of the designated levels outlined in *Leveled Books K–8: Matching Texts to*

Readers for Effective Teaching (Fountas & Pinnell, 2006a). At every level, with both fiction and nonfiction, the *Fountas & Pinnell Benchmark Assessment System* books are distinguished by their writing quality, compelling content, use of universal concepts, and visually strong illustrations. Text length is appropriate for grade level. In *System 1*, used for grades K–2, text levels A–N are 16 pages in length. In *System 2*, used for grades 3–8, text levels L–Z are four pages in length. Each *System* provides the teacher, evaluator, or clinician an appropriate measure to assess a student's reading accuracy, fluency, and comprehension level for informing instruction.

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). As part of the editorial development process, Fountas and Pinnell selected two separate teams of classroom teachers, one team to vet books for *System 1*, and the other team to vet books for *System 2*. These educators were chosen based on their experience in teaching with Fountas and Pinnell leveled books. These leveling teams met on three 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

Field testing included a total of 497 students spanning grades K–8. Field testing of *System 1* included 252 students and *System 2* included 245 students. School sites from which these students were drawn were socioeconomically, ethnically, and geographically diverse. The research goal was to identify “typical students.” Accordingly, students were selected on the basis of their ability to read and understand texts that were written approximately at grade level or above. Participants were also proficient speakers of English. Each field test examiner determined an individual student's eligibility after discussing his or her reading profile with their respective teachers.

Field-Test Examiners

Thirteen field-test examiners were selected. All field-test examiners were educators who had extensive training in administering Reading Running Records (Clay, 2002) and in using other forms of benchmark assessments to assess students' reading levels. Field-test examiners were not affiliated with the field sites and therefore could be objective in both identifying students and in administering assessments. Prior to the

beginning of the field testing, a two-day intensive training session led by the program’s authors, Irene Fountas and Gay Su Pinnell, guided the field-test examiners in the formative evaluation’s protocols and procedures.

Contexts

A total of 22 different schools participated in field testing of either *System 1* or *System 2* (some schools participated in both field tests). Field testing took place across the following geographic regions of the United States:

- Boston Metropolitan area 1 examiner; 1 school
- Providence, Rhode Island 1 examiner; 2 schools
- Houston Metropolitan area 2 examiners; 5 schools
- Los Angeles area 4 examiners; 6 schools
- Columbus, OH, area 3 examiners; 5 schools
- Orlando, FL, area 2 examiners; 3 schools

A second round of field testing for *System 2* was conducted in four of the six original geographic locations (Ohio, California, Texas, and Florida).

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 using federal guidelines for categorizing low-, middle-, and high-SES schools. Therefore, students in Phase I and Phase II represented a cross-section of the major regions of the U.S. and diverse socioeconomic levels (see Figure 2).

SCHOOL FIELD SITES						
	California	Florida	Massachusetts	Ohio	Rhode Island	Texas
Number of school sites (elementary and middle school)	6	3	1	5	2	5
Average percentage of students receiving free or reduced-price lunch or economically disadvantaged	64.2%	45.3%	71%	28.4%	51.5%	46.4%

Figure 2

A broad range of students from diverse ethnic backgrounds participated in the study. The chart below shows the average percentage by ethnicity from the school field sites from each state.

SCHOOL FIELD SITES							
	CA	FL	MA	OH	RI	TX	Overall Average
African American	7.3%	22.3%	41%	1.8%	18.5%	34%	20.8%
Asian American & Pacific Islander	5.5%	5.3%	23%	0%	3%	17%	9%
Hispanic/Latino	74%	23%	6%	0%	24%	30%	26%
White	11.7%	45.7%	29%	92.6%	54%	19%	42%
Multiracial/Other	1.5%	3.7%	1%	5.6%	0.5%	0%	2.1%

Figure 3

PROCEDURES AND MATERIALS

Phase I of the Formative Evaluation

Phase I 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 books were tested in the following sequence:

1. Fiction, *System 1*, levels A–N (grades K–2)
2. Nonfiction, *System 1*, levels A–N (grades K–2)
3. Fiction, *System 2*, levels L–Z (grades 3–8)
4. Nonfiction, *System 2*, levels L–Z (grades 3–8)

Procedures for Assessment Administration

Reading data for every student using both fiction and nonfiction books was gathered systematically through a formative evaluation design protocol. After an intensive training session, the field test examiners began working individually at selected school sites during the last quarter of 2006. By conferring with classroom teachers at each site, field test examiners identified eligible students who met the criteria for inclusion in the study (i.e., students who were considered to be “typical” readers according to grade level norms). Below is a list of protocols and procedures followed by each field-test examiner:

Selecting a starting point for reading

A Where to Start word list was developed by the program’s authors to assist field-test examiners in quickly placing a student at his or her appropriate reading level. This word list was

administered to all eligible students in their classrooms. Using this as a starting point, the field test examiners readily determined which book they should ask the student to read first.

Determining a decoding instructional reading level

Next, field-test examiners assessed each student's ability to read and comprehend 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). Accuracy of reading guidelines, consistent with Fountas and Pinnell's framework (2006b), is as follows: independent level (95–100 percent accuracy); instructional level (90–94 percent accuracy), and hard level (below 90 percent accuracy).¹

Determining comprehension instructional reading level

Once field-test examiners determined a student's instructional reading level, they engaged in a comprehension conversation about that particular book. If students were unresponsive or gave an incomplete response, field-test examiners prompted them according to a predetermined set of questions. Next, field-test examiners rated students' understanding of a text using the Fountas and Pinnell comprehension guidelines (2001, pp. 323–24). The focal areas listed below were rated on a scale from 0–3:

- a. Thinking within the text
- b. Thinking beyond the text
- c. Thinking about the 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 students should read along at a reasonable pace when reading at their instructional level. Consistent with Fountas and Pinnell's fluency assessment guidelines (2001, pp. 491–92), which draw upon the National Assessment of Educational Progress (NAEP) Integrated Reading Performance Record Oral Reading Fluency Scale, the field-test examiners rated readers' fluency across the following three dimensions. (Note that this scale is applicable only for students in grades 3–6.)

1. Readers phrase, or group words, through intonation, stress, and pauses. They emphasize the beginnings and endings of phrases by the rising and falling of pitch or by pausing.
2. Students adhere to the author's syntax or sentence structure, reflecting their comprehension.
3. Readers are expressive; their reading reflects feeling, anticipation, and character development.

Determining the corresponding readability between fiction and nonfiction books

Finally, the field-test examiners repeated the process described above, with the same students, using nonfiction books. Given that students' reading levels had been established, the field-test examiners did not need to re-administer the word list test. Field-test examiners began the session reading nonfiction books at the students' instructional levels. They concluded the session when all three sequential levels of a student's reading had been ascertained: independent, instructional, and hard.

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 or students' interest in subject matter, motivation, need for warm-up time, 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 either by phone or in person throughout the testing process.

Schedule of Assessment Administration

The field-test examiners worked on a somewhat staggered schedule. This allowed them to refine the day-to-day practical aspects of the research as needed and to immediately replace any books that tested out of order.

In general, the schedule flowed as indicated in Figure 4.

¹ Based on feedback of the field testing, new accuracy criteria were established for Benchmark System 2 (levels L–Z) establishing a finer gradient reflection of students' reading in grades 3 through 8. A discussion of the change and the new accuracy criteria are provided in this report's section "Formative Program Development" on pages 5–6.

THREE-DAY SCHEDULE FOR FIELD-TEST EXAMINERS	
DAY 1: Fiction Texts	
<ul style="list-style-type: none"> • Administer word test • Ascertain instructional, independent, and hard levels with <i>fiction</i> texts • Work with approximately 12 students 	
DAY 2: Nonfiction Texts	
<ul style="list-style-type: none"> • Work with same students • Ascertain instructional, independent, and hard levels with <i>nonfiction</i> texts • Each student should read at least three books 	
DAY 3: Creating Additional Data	
<ul style="list-style-type: none"> • Return to classrooms to obtain data missing because of student absences, school field trips, assemblies, scheduling conflicts with district programs, etc. • Work with approximately six additional students reading <i>fiction</i> and <i>nonfiction</i> books 	

Figure 4

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. Documentation was completed on recording forms (see Appendix A for sample) to capture students’ reading accuracy, fluency, and comprehension scores as well as other data. This included recording the book titles—both fiction and nonfiction—that had been read.

Research Debriefings

On a daily basis, the field-test examiners analyzed new data collected in the field and reported back to the research project manager in debriefings by phone and email. These daily debriefings provided an opportunity to take immediate action on revising texts (if a particular book tested poorly, for example).

FORMATIVE PROGRAM DEVELOPMENT

With a formative evaluation process, data analysis was ongoing as well as recursive. Ongoing field data gathered in authentic contexts provided immediate information for adjustments and revisions in the program. As described previously, the research project manager debriefed field-test examiners at the end of each day’s field testing data collection. Based on these data, the research project manager identified patterns both within individual books and across books using the following four categories of text evaluation:

1. Texts that were completely on-target
2. Texts that required minor revisions

3. Texts that required substantive revisions
4. Texts that needed to be replaced altogether.

In December 2006 and January 2007, based on the *System 1* field test results, changes were made in the leveling of the texts. For example, the level C fiction book *Big Lizard, Little Lizard* was replaced by *Socks*, and the nonfiction text *Earthquake* was changed from level U to level V. Drawing upon students’ reading data, the research project manager made specific recommendations for the gradient of difficulty represented across several dimensions. One example was modifying the texts to increase their appropriateness for their designated level. These changes included simplifying the specialized vocabulary words in some nonfiction texts or recasting sentences in a particular text to make them either more or less complex. At one point, two books were replaced with more appropriate books.

At the beginning of January 2007, based on the *System 2* field test results, major changes were made to specific texts. These changes included modifying vocabulary, sentence complexity, or text selection.

After another round of field testing in January and February 2007, field-test examiners discovered a need to revise comprehension criteria because they found some students were able to decode increasingly difficult texts without the corresponding comprehension. Students had an independent level and a hard level, but no instructional level. This was especially prevalent with upper-elementary students. This ability to decode is a not an uncommon occurrence in any reading assessment program. However, responding to this concern, the program

BENCHMARK CRITERIA FOR LEVELS A-K				
Accuracy	Comprehension			
	Excellent 6-7	Satisfactory 5	Limited 4	Unsatisfactory 0-3
95%–100%	Independent	Independent	Instructional	Hard
90%–94%	Instructional	Instructional	Hard	Hard
Below 90%	Hard	Hard	Hard	Hard

BENCHMARK CRITERIA FOR LEVELS L-Z				
Accuracy	Comprehension			
	Excellent 6-7	Satisfactory 5	Limited 4	Unsatisfactory 0-3
98%–100%	Independent	Independent	Instructional	Hard
95%–97%	Instructional	Instructional	Hard	Hard
Below 95%	Hard	Hard	Hard	Hard

Figure 5

developers recognized the need for educators to establish an instructional level as a stopping place. As a result, the *Fountas & Pinnell Benchmark Assessment System* includes new parameters (see Figure 5) linking accuracy and comprehension with independent, instructional, and hard reading levels. The accuracy levels were changed for *System 2* levels L–Z, with the same criteria for comprehension. Developing new parameters is an innovative step in assisting educators with a more finely grained reflection of a student’s decoding coupled with an appropriate understanding of his or her text reading.

DATA ANALYSIS AND RESULTS: RELIABILITY AND VALIDITY

When the field testing was completed, an independent team of three research specialists was brought in to analyze the data. This team consisted of researchers experienced in quantitative data analysis as well as research design, methods, and data collection.

Phase I of the Formative Evaluation

Phase I 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 horizontal readability between fiction and nonfiction books. The results of Phase I 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

- How reliable is the *Fountas & Pinnell Benchmark Assessment System*? That is, how consistent and stable is the information derived from the reading books?
- Does each book of the *Fountas & Pinnell Benchmark Assessment System* consistently occupy the same position on the gradient of readability, based on multiple readings by age-appropriate students? That is, does each book, from A–Z, represent a degree of increased difficulty that is consistent with other Fountas and Pinnell leveled texts?

The findings, obtained from field testing conducted in varied geographic regions throughout the country, demonstrate that relative to the text gradient, the *Fountas & Pinnell Benchmark Assessment System* books get progressively more difficult as the levels progress vertically from A–Z.

Section 1. Data Analysis of Vertical Text Gradient

All students with complete data were included in the analysis. Students for whom an instructional level had been identified and

who had also been tested on the books immediately preceding and succeeding the instructional level were included. Students that had an instructional level with test information for one level higher or lower than the immediate and/or subsequent levels were also included. Students were not included if they had not tested at an instructional level, or where data for the preceding and/or succeeding levels were not available.

Section 1. Findings of Vertical Text Gradient

There were two ways in which students read the text gradient. Students read leveled texts (1) in sequential and hierarchical progression or (2) with some degree of variation. The following describes each.

i. Sequential and Hierarchical Progression from Lower to Higher Levels of Text Difficulty

The students’ reading progression from lower levels on the A–Z Text Gradient to higher levels was sequential and hierarchical. That is, 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 the following chart (Figure 6).

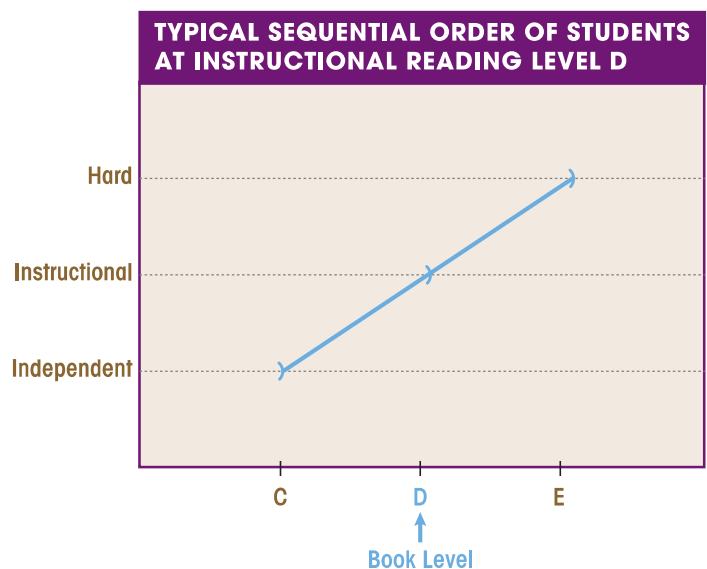


Figure 6

For *System 1* (grades K–2), 60.4% of the students read the fiction books and 53.8% read the nonfiction texts in sequential and hierarchical order. For *System 2* (grades 3–8), 80.3% of the students read the fiction texts and 75.4% read the nonfiction texts in sequential and hierarchical order. The following table (Figure 7) depicts the results.

VERTICAL TEXT GRADIENT SEQUENTIAL AND HIERARCHICAL PROGRESSION FROM LOWER TO HIGHER LEVELS OF DIFFICULTY		
	Benchmark System 1 (Levels A–N)	Benchmark System 2 (Levels L–Z)
Fiction	60.4%	80.3%
Nonfiction	53.8%	75.4%

Figure 7

ii. Variations in the Sequential and Hierarchical Progression from Lower to Higher Levels of Difficulty

In the previous section, the students’ reading progression from lower to higher levels of difficulty was described as occurring sequentially from one level to the next. However, the students’ progressions through the levels included some variations that are discussed below.

(1) Level immediately preceding instructional level was not easier

Some students’ progression from the instructional level to the subsequent level was more difficult. This indicated a sequential, hierarchical pattern of increased difficulty; however, the book preceding the instructional level was not independent or easier. The instructional level was therefore the same degree of difficulty (or easier than) the immediately preceding level. For example, when level D was the instructional level, level C was also at the instructional level, and both were more difficult than the preceding level B, but less difficult than the subsequent level E. When analyzing the reading scores of the books within one level of the preceding book, the books became easier, indicating a sequential and hierarchical pattern, as illustrated in Figure 8.

(2) Level immediately succeeding instructional level was not more difficult

In other cases of divergent sequential ordering, the students’ progression between the instructional level and the preceding level was easier. These findings indicated a sequential, hierarchical pattern of increased difficulty. However, the book succeeding the instructional level was not more difficult. The instructional level was therefore the same degree of difficulty as (or harder than) the immediately succeeding level. For example, when level D was the instructional level, the preceding level C was easier, but the subsequent level E was also at the instructional level; level F was more difficult than both levels D and E. Therefore, when analyzing the reading scores within one level of the succeeding book, the books became harder, indicating a sequential and hierarchical pattern, as illustrated in Figure 9.

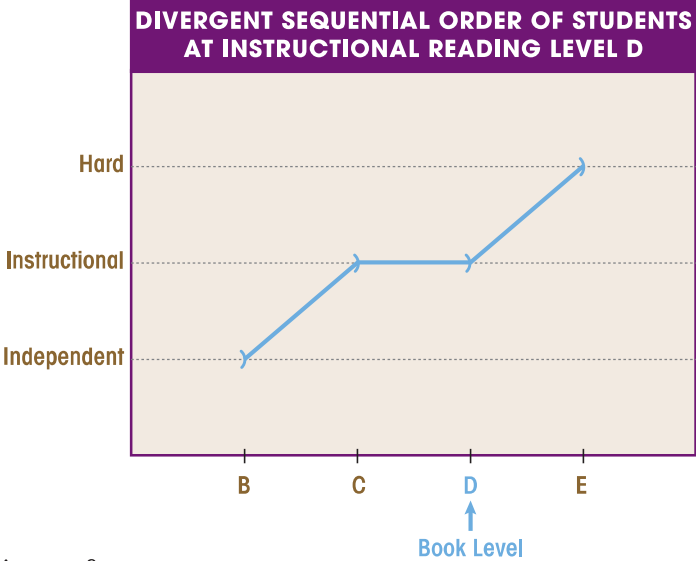


Figure 8

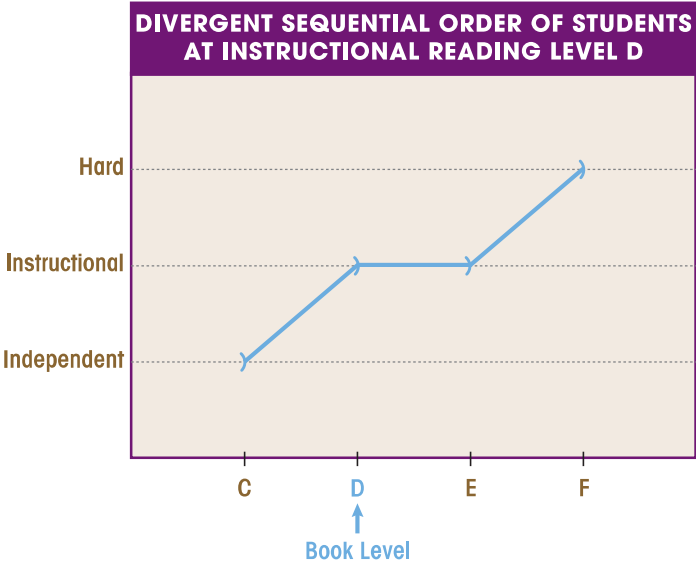


Figure 9

Section 1. Results of the Vertical Text Gradient

The findings in section 1 regarding the vertical text gradient indicated that texts became more difficult as a reader progressed through them in sequence. When the sequence was expanded to include one level below the preceding reading level or one level above the succeeding reading level, the gradient percentage increased, reflecting a stronger vertical text gradient. For *System 1* (grades K–2), 81.1% of the students now read the fiction texts in a divergent, but sequential and hierarchical order, and 80.4% now read the nonfiction books in that order. For *System 2* (grades 3–8), 95.8% of the students now read the fiction texts in sequential and hierarchical order, and 84.2% read the nonfiction texts in that order. The following table (Figure 10) depicts the results.

VERTICAL TEXT GRADIENT DIVERGENT, BUT SEQUENTIAL AND HIERARCHICAL PROGRESSION FROM LOWER TO HIGHER LEVELS OF DIFFICULTY		
	System 1 (Levels A–N)	System 2 (Levels L–Z)
Fiction	81.1%	95.8%
Nonfiction	80.4%	84.2%

Figure 10

The following charts (Figures 11 through 16) depict the percentage of students who read in a sequential and hierarchical order from lower to higher levels of difficulty when the sequence was expanded to include one level above or below the levels preceding and succeeding the targeted reading level.

Benchmark Assessment System 1: Fiction and Nonfiction

The first two charts (Figures 11 and 12) represent the progress of students reading the *System 1* fiction and nonfiction books (levels A–N) in the sequential order when the sequence was expanded to include one level above or below the levels preceding and succeeding the targeted reading level.

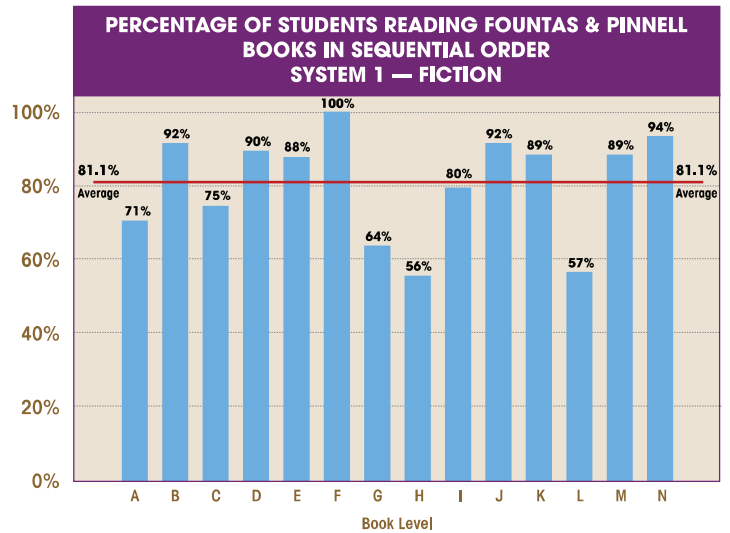


Figure 11

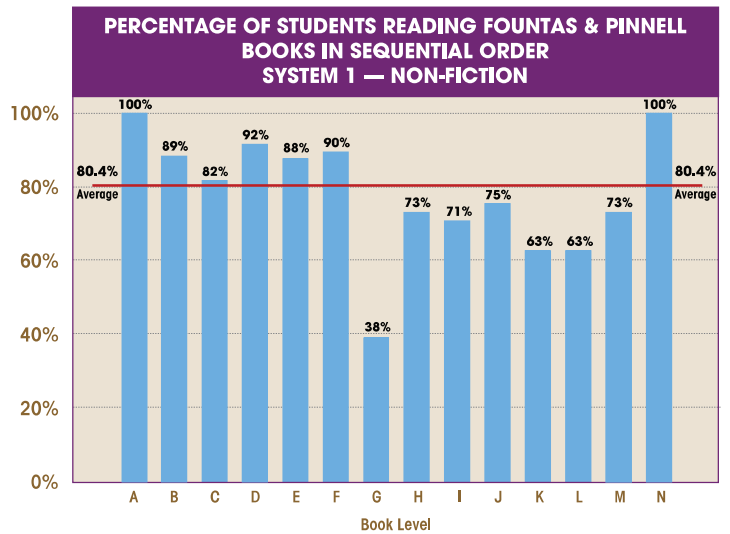


Figure 12

Benchmark Assessment System 2: Fiction and Nonfiction

The next two charts (Figures 13 and 14) represent the progress of students reading *System 2* fiction and nonfiction books (levels L–Z) in the sequential order when the sequence was expanded to include one level above or below the levels preceding and succeeding the targeted reading level.

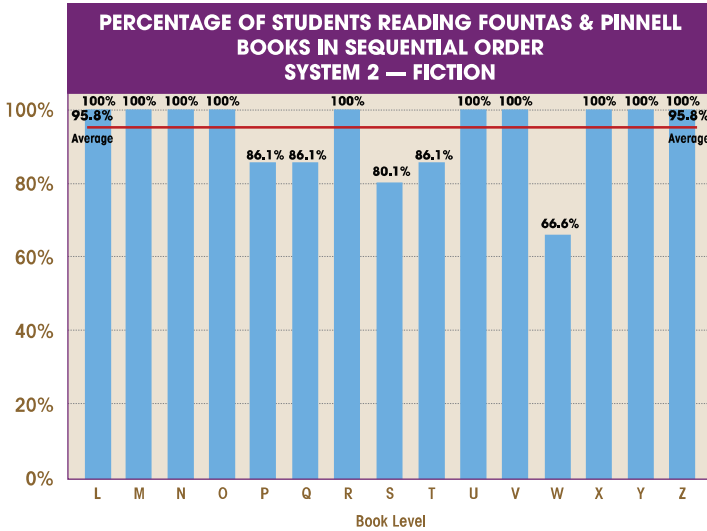


Figure 13

Benchmark Assessment Systems 1 & 2: Fiction and Nonfiction by Grade Level

The final two charts (Figures 15 and 16) represent the progress of students reading the *Fountas & Pinnell Benchmark Assessment System 1 and 2* fiction and nonfiction books (levels A–Z) by grade level, in the sequential order when the sequence was expanded to include one level above or below the levels preceding and succeeding the targeted reading level.

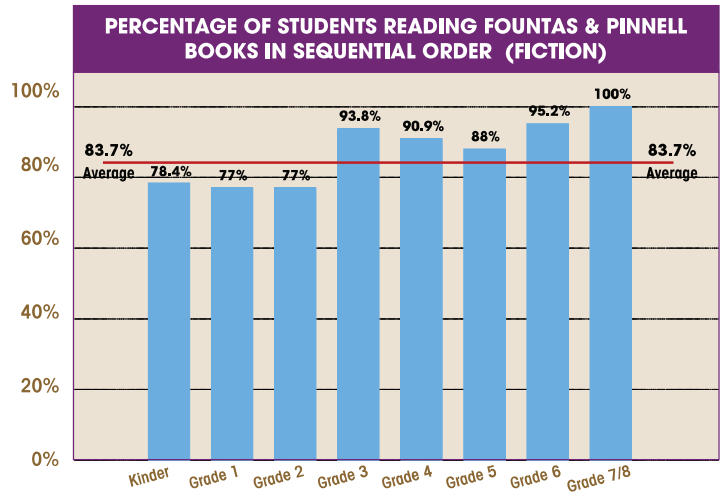


Figure 15

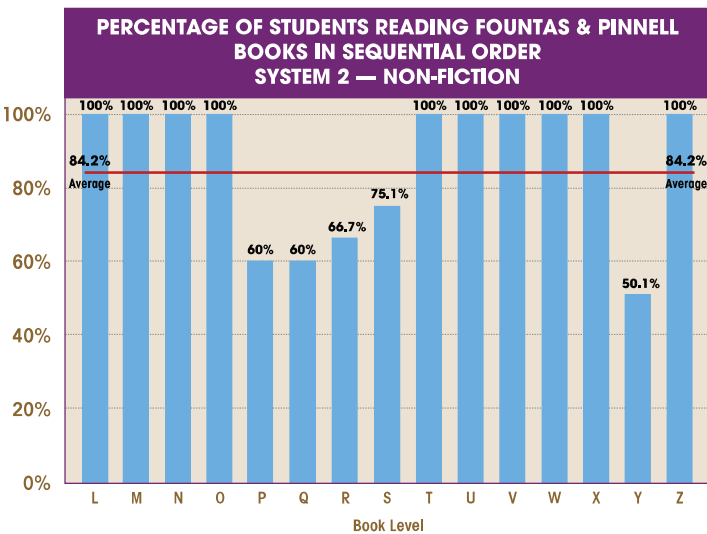


Figure 14

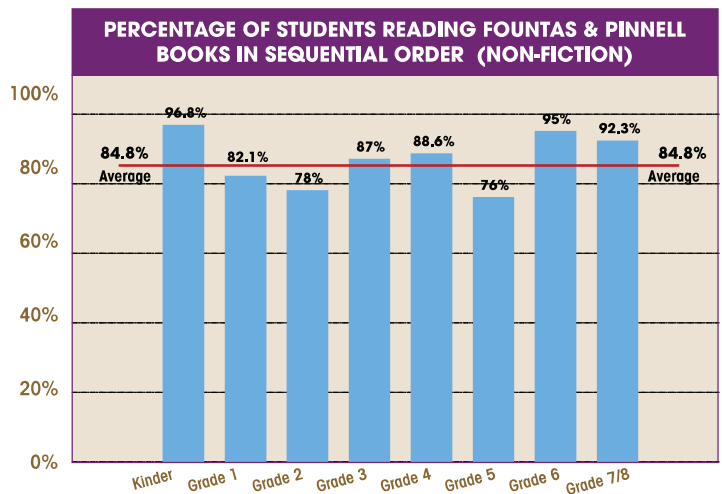


Figure 16

Section 2. Horizontal Consistency Between Fiction and Nonfiction Texts

Research Question 2

- To what extent are the gradients of difficulty for *fiction* and *nonfiction* books aligned within the *Fountas & Pinnell Benchmark Assessment System*? Do fiction and nonfiction books represent similar levels of difficulty within similar levels of reading?

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. In other words, are the fiction and nonfiction books consistent, representing a similar level of difficulty at each level, A–Z, on the Text Gradient? 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 varied geographic regions throughout the country, indicate that relative to the consistency of the difficulty of the fiction and nonfiction texts, the books are written at similar levels of difficulty at each level of the A–Z text gradient.

Section 2. Data Analysis of Horizontal Text Consistency

All students with complete data were included in the analysis. Only students that had an instructional level in both fiction and nonfiction were included.

Section 2. Results of Horizontal Text Consistency

A preponderance of students read the text gradient in two ways. Students read fiction and nonfiction texts at the same level on the A–Z Text Gradient or they read fiction and nonfiction texts at similar levels of difficulty on the A–Z Text Gradient. The following describes each.

i. Fiction and Nonfiction Texts Represent Same Level of Text Difficulty

The students’ developmental reading level is the *same* for fiction and nonfiction on the A–Z 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 *System 1* (grades K–2), 43.4% of the students read at the same level in fiction and nonfiction. For *Benchmark System 2* (grades 3–8), 26.1% of the students read at the same level in fiction and nonfiction. Figure 17 depicts these results.

There are many factors underlying the 26.1% correspondence for students in grades 3–8 reading fiction and nonfiction texts at the same level. One explanatory factor is that as readers progress through the grade levels, their mastery of content knowledge plays an increasingly larger and complex role

HORIZONTAL TEXT GRADIENT: STUDENTS READING AT THE SAME LEVEL OF TEXT DIFFICULTY ON FICTION AND NONFICTION TEXTS		
	System 1 (Levels A–N)	System 2 (Levels L–Z)
Fiction–Nonfiction	43.4%	26.1%

Figure 17

HORIZONTAL TEXT GRADIENT: STUDENTS READING WITHIN ONE LEVEL OF TEXT DIFFICULTY ON THE FICTION AND NONFICTION TEXTS		
	System 1 (Levels A–N)	System 2 (Levels L–Z)
Fiction–Nonfiction	76.2%	69.2%

Figure 18

in reading comprehension. In other words, it is difficult to predict, given a classroom’s instructional focus and students’ background knowledge, what a student might “typically” know.

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–Z Text Gradient. For example, a student reading at an instructional level D on a fiction text would read on an instructional level at the nonfiction text of the preceding level (level C) or succeeding level (level E).

When the analysis was expanded to include one level above or below the instructional level on the fiction text for the nonfiction text, the gradient percentage increased to reflect a stronger horizontal text gradient. For *System 1* (grades K–2), 76.2% of the students read at an instructional level on the nonfiction text within one level of difficulty on the fiction text. For *System 2* (grades 3–8), 69.2% of the students read the fiction and nonfiction texts within one level of difficulty. Figure 18 depicts these results.

Fountas and Pinnell Benchmark Assessment Systems 1 & 2

For *System 1* and *System 2* combined, 75.8% of the students read the fiction and nonfiction texts within one level of difficulty. The following chart (Figure 19) shows by grade level, the percentage of students reading the fiction and nonfiction texts within one level of text difficulty.

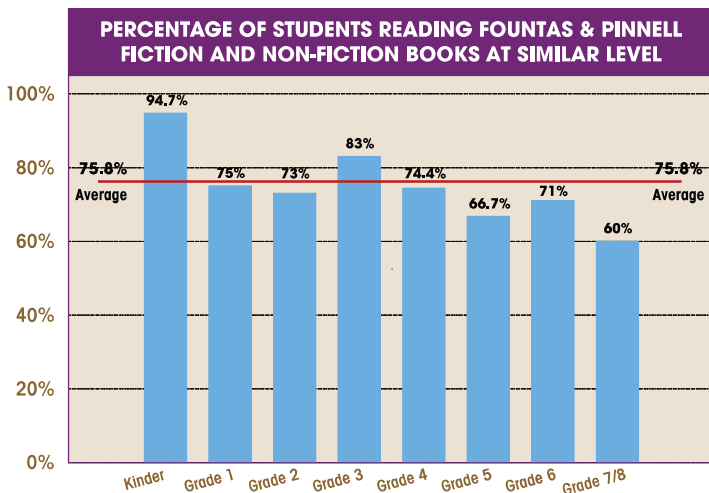


Figure 19

Phase II of the Formative Evaluation

Phase II of the formative evaluation examined research question 3 related to the reliability and validity of the *Fountas & Pinnell Benchmark Assessment System*.

Research Question 3

- How reliable is the *Fountas & Pinnell Benchmark Assessment System*? That is, how consistent and stable is the information derived from the reading books?
- To what extent is the *Fountas & Pinnell Benchmark Assessment System* associated with other established reading assessments?

RELIABILITY

Reliability addresses the consistency of scores of an assessment, in this case the *Fountas & Pinnell Benchmark Assessment System*. 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 the *Fountas & Pinnell Benchmark Assessment System*, 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.

In general, test-retest results should exhibit a reliability coefficient of at least .85 for the assessment's information to be considered stable, consistent, and dependable. As the test-retest results depicted in Figure 20 demonstrate, the *Fountas & Pinnell Benchmark Assessment System* is a reliable reading assessment.

Book Series A–N	.93
Book Series L–Z	.94
All Books (A–Z)	.97

Figure 20

CONVERGENT VALIDITY

The validity of a 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 an assessment's test scores and the scores from other instruments that measure similar variables. Therefore, the assessment outcomes from the *Fountas & Pinnell* texts should be related with other tests that assess reading.

- For *System 1* correlation to texts used for assessments in Reading Recovery[®].
- For *System 2* correlation to Slosson Oral Reading Test—Revised (SORT-R3) and the Degrees of Reading Power[®].

METHODS

Three teams of field-test examiners (field-test examiners worked individually) followed the procedure described previously in phase I to determine students' independent, instructional, and hard levels of reading proficiency at grade levels K–8 on the *Fountas & Pinnell Benchmark Assessment System*. Next, field-test examiners administered the Slosson Oral Reading Test—Revised (SORT-R3) to students in grades 3 through 8. Then, field-test examiners administered either Reading Running Records with texts used for Reading Recovery[®] assessments or the Degrees of Reading Power[®] (DRP), according to whether their instructional level aligned with *System 1* or *2*. As in Phase I, field-test examiners systematically maintained data records and participated in daily debriefings with the research project manager.

DESCRIPTION OF OTHER ASSESSMENTS

The *Reading Recovery[®] Observation Survey Text Reading Level*. The Observation Survey consists of six literacy tasks; one is the Text Reading Level. This task records, by using a running record of a student's reading, the accuracy and process the child employs

when reading. Increasingly difficult texts are used to ascertain his or her appropriate reading level. In recent studies, the Text Reading Level was correlated with other standardized, norm-referenced tests. These include the Iowa Test of Basic Skills (.764 correlation) (Gómez-Bellengé, Rodgers, Wang, & Schulz, 2005), the Gates-MacGinitie Reading Test, and the Woodcock Reading Mastery Test (Gómez-Bellengé & Thompson, 2005).

The Degrees of Reading Power® (DRP) is a norm-referenced assessment made up of nonfiction text passages formatted using a cloze technique. That is, selected words are omitted from the text and the student selects a word from among multiple choices. DRP measures basic comprehension. The assessment measures where to place a reader on a range of texts. Based on a student's assessed frustration, instructional, or independent level, placement is determined for appropriate reading materials.

The Slosson Oral Reading Test-Revised (SORT-R3) is a list of 200 words in increasing order of difficulty administered individually to students. Words are grouped together in ten lists of 20 words and each list corresponds to a specific grade level. Although it does not measure comprehension—it measures students' oral word calling—the assessment assists educators in providing placement on a child's approximate reading level.

RESULTS OF CONVERGENT VALIDITY

Convergent Validity with Reading Recovery Assessment Texts

The table (Figure 21) shows a strong relationship between the reading accuracy rates on *System 1* (levels A–N) fiction (correlation of .94) and nonfiction (correlation of .93), and reading accuracy rates on texts used for assessments in Reading Recovery®. In other words, the performance on the *System 1* texts is strongly indicative of performance on Reading Recovery® assessment books. This is an important finding because the Reading Recovery Text Level assessments, like the *Fountas & Pinnell Benchmark Assessment System*, assess decoding, fluency, vocabulary, and comprehension. In addition, Reading Recovery® was recognized in March 2007 by the U.S. Department of Education as an effective and scientifically based reading program (see: What Works Clearinghouse, 2007). These results reinforce the validity of the *Fountas & Pinnell Benchmark Assessment System 1* program.

Convergent Validity with Slosson Word Test

Another aspect of Phase II of the formative evaluation examined the relationship between the *System 2* fiction and nonfiction books (levels L–Z) and the Slosson Word Test. The results in Figure 21 indicate that performance on the *System 2* fiction texts (correlation of .69) and nonfiction texts (correlation of .62) is moderately indicative of performance on Slosson word reading. The Slosson Word Test measures students' oral word calling and provides

RELATIONSHIP BETWEEN FOUNTAS & PINNELL BENCHMARK ASSESSMENT SYSTEM INSTRUCTIONAL READING LEVELS AND OTHER MEASURES OF INSTRUCTIONAL READING			
<i>Fountas & Pinnell Benchmark Assessment System</i>	Other Measures	Benchmark Assessment System Fiction Books	Benchmark Assessment System Nonfiction Books
Text Levels A–N	Reading Recovery® Text Level Assessment Books	.94	.93
Text Levels L–Z	Slosson Word Test	.69	.62
Text Levels L–Z	DRP®	.44	.42

Figure 21

approximate placement of a child's reading level. These results indicate that the *Benchmark System 2* texts are moderately indicative of the Slosson measure of word reading. It should be emphasized, however, that the *Fountas & Pinnell Benchmark Assessment System* is more than a word reading measure.

When the *Fountas & Pinnell Benchmark Assessment System* grade levels were compared with Slosson grade levels for fiction and nonfiction books, students generally scored higher on the Slosson than they did with the *Fountas and Pinnell Benchmark Assessment System* texts for grades 3–6. However, this pattern did not occur in grades 7 and 8. Because Slosson measures only isolated word reading, it can be expected that students might score higher when compared with the *Fountas & Pinnell Benchmark Assessment System* in which a student's score is based on comprehensive reading of complete books.

Convergent Validity with Degrees of Reading Power

A final study looked at the relationship between the *System 2* (Levels L–Z) books and the DRP® text passage reading. The DRP is made up of nonfiction text passages using a cloze technique and measures where to place a reader on a range of texts. The *Benchmark System 2* fiction books (correlation of .44) and nonfiction books (correlation of .42) were moderately related with performance on DRP. These results (Figure 21) therefore show that the *Benchmark System 2* texts are moderately indicative of cloze text passages. It should be noted, however, that the *Fountas & Pinnell Benchmark Assessment System* includes a reading comprehension dimension, through question prompts, in the context of complete books, while the DRP® measures comprehension as the degree to which the reader accurately predicts the words missing in short passages through multiple word choices.

SUMMARY AND CONCLUSION

Summary of Findings

Research question 1 asked whether each book from levels A–Z 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 *Fountas & Pinnell Benchmark Assessment System* books get progressively more difficult as the levels progress vertically from A–Z.
- For *System 1* (grades K–2), 81.1% of the students read the fiction texts and 80% read the nonfiction books in a divergent but sequential and hierarchical order. For *System 2* (grades 3–8), 95.8% of the students read the fiction texts and 84.2% read the nonfiction texts in a divergent but 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 *Fountas & Pinnell Benchmark Assessment System* series. That is, do fiction and nonfiction books represent similar levels of difficulty within similar levels of reading?

- For *System 1* (grades K–2), 76.2% of the students read at similar levels in fiction and nonfiction within one level of text difficulty.
- For *System 2* (grades 3–8), 69.2% of the students read at similar levels in fiction and nonfiction within one level of text difficulty.

Research question 3 addressed the reliability and validity of the *Fountas & Pinnell Benchmark Assessment System* with other assessment measures.

- There was a strong association between the *System 1* (levels A–N) fiction texts (correlation of .94) and nonfiction texts (correlation of .93) and Reading Recovery® Text Level Assessments. This is an important finding, since Reading Recovery® was recently recognized by the U.S. Department of Education as an effective and scientifically based reading program.
- The results indicate that performance on the *System 2* fiction texts (correlation of .69) and nonfiction texts (correlation of .62) is moderately indicative of performance on Slosson word reading. Again, it needs to be emphasized that the *Fountas & Pinnell Benchmark Assessment System 2* is more than a word reading measure, since it adds a reading comprehension dimension.
- The *System 2* fiction texts (correlation of .44) and nonfiction texts (correlation of .42) were moderately indicative of performance on DRP® word reading.

Conclusion

After two and a half years of editorial development, field testing, and independent data analysis, the *Fountas & Pinnell Benchmark Assessment System* texts were demonstrated to be both reliable and valid measures for assessing students' reading levels.

The final report was compiled by an outside team of three independent researchers who analyzed the data gathered from the formative evaluation of the Fountas & Pinnell Benchmark Assessment Systems 1 and 2. Two research team members were former school literacy coaches and Reading Recovery educators. All data analysts had backgrounds in literacy research studies using quantitative and qualitative methods and analysis. The final report incorporated the initial formative evaluation design, methods, and collected data.

Appendix A: Field Test Examiner Recording Form Samples

REFERENCES

- Clay, M. M. (2002). *An observation survey of early literacy achievement* (2nd ed.). Portsmouth, NH: Heinemann.
- Fountas, I., & Pinnell, G. S. (2001). *Guiding readers and writers*. Portsmouth, NH: Heinemann.
- Fountas, I., & Pinnell, G. S. (2006a). *Leveled books K-8: Matching texts to readers for effective teaching*. Portsmouth, NH: Heinemann.
- Fountas, I., & Pinnell, G. S. (2006b). *Teaching for comprehension and fluency*. Portsmouth, NH: Heinemann.
- Gómez-Bellengé, F., & Thompson, J. R. (2005). *U.S. Norms for Tasks of An Observation Survey of Early Literacy Achievement*. Reading Recovery/Descubriendo la Lectura, National Data Evaluation Center. The Ohio State University College of Education, School of Teaching and Learning, Reading Recovery. NDEC 2005-02. Retrieved 6/20/07 from Reading Recovery/Descubriendo la Lectura, National Data Evaluation Center website, Publications - Methodology: <http://www.ndec.us/Documentation.asp>
- Gómez-Bellengé, F., Rodgers, E., Wang, C., & Schulz, M. (2005). *Examination of the Validity of the Observation Survey with a Comparison to ITBS*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec. Retrieved 6/20/07 from Reading Recovery/Descubriendo la Lectura, National Data Evaluation Center website, Publications - Presentations: <http://www.ndec.us/Documentation.asp>
- Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud: Data from NAEP's Integrated Reading Performance Record (IRPR) at Grade 4*. Report No. 23-FR-04. Prepared by Educational Testing Service under contract with the National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education, p. 15.
- What Works Clearinghouse Intervention Reports. (2007, March 19). *Intervention: Reading Recovery*. United States Department of Education, Institute of Education Sciences. Retrieved June 20, 2007 from: www.whatworks.ed.gov/PDF/Intervention/WWC_Reading_Recovery_031907.html

APPENDIX A: FIELD-TEST EXAMINER RECORDING FORM SAMPLES

Fun at the Park, Level A, Nonfiction, RW: 24

Field Test Examiner: _____ Location: _____ Date: _____

Student: _____ Grade: _____

TEXT READING SUMMARY
Fun at the Park, Level A, Nonfiction

Oral Reading Summary

Scores	Comments on Reading Behaviors
Accuracy: Independent _____ %	
Instructional _____ %	
Hard _____ %	
Errors: _____	
Self-Corrections _____	
Self-Correction Ratio: I: _____	Comments on Fluency (Consider pausing, phrasing, stress, intonation, rate, and integration.)
(E + SC) SC _____	
Fluency	
Oral Reading Rate: _____ WPM	
Fluency: _____ 3 2 1 0	

Comprehension Summary

Scores	Comments on Comprehension Conversation
Thinking Within Text: _____ /3	
Thinking Beyond Text: _____ /3	
Additional Understandings: _____ /1	
Total Score: _____ /7	

Comments on Response to Reading (optional Writing/Drawing activity)

Field Test Examiner's Comments

Fountas & Pinnell Benchmark Assessment System 1 1

Fun at the Park, Level A, Nonfiction, RW: 24

Student: _____ Date: _____

A. ORAL READING
Fun at the Park, Level A, Nonfiction
Read the book title to child before the introduction.

Introduction: These children are all having fun at the park. Read to find out all of the things they are doing to have fun at the park. You may point to each word as you read.

Start Time: _____ End Time: _____ Total Time: _____ Oral Reading Rate: _____ WPM WPM = 24 (RW) x 60 (sec) = 1440
sec. child reads

Page	Text	E	SC	Information Used	
				MSV	MSV
2	I can ride.				
4	I can kick.				
6	I can catch.				
8	I can jump.				
10	I can swing.				
12	I can slide.				
14	I can run.				
16	I can hide.				
Totals					

Fountas & Pinnell Benchmark Assessment System 1 2

Fun at the Park, Level A, Nonfiction, RW: 24

(cont.)

Accuracy

Errors	3 or more	2	1	0
%	Below 90%	92%	96%	100%
	Hard	Instructional	Independent	

Fluency Rubric
Assess child's ability to read the text fluently.

3 2 1 0

Rubric Key:

0 = no phrasing or expression
1 = minimal phrasing or expression
2 = some phrasing or expression
3 = mostly phrased and expressive reading

Fountas & Pinnell Benchmark Assessment System 1 3

APPENDIX A: FIELD-TEST EXAMINER RECORDING FORM SAMPLES (CONTINUED)

Fun at the Park, Level A, Nonfiction, RW: 24

B. COMPREHENSION CONVERSATION
 Instructions: Let's talk about what you learned in this book.
 Have a conversation with the child about the text. If the child provides evidence of key understandings, circle the appropriate number. Use the prompts as needed to stimulate discussion of key understandings the child did not talk about.

	Rubric Key: 0 = no understanding 1 = minimal understanding 2 = partial understanding 3 = complete understanding	
--	---	--

Key Understandings	Prompts	Rubric
Within the Text The boy and his dad are at the park. There are lots of things to do at the park. The boy is having fun doing <u>(activities)</u> at the park. Note any additional understandings:	Talk about what the boy and his dad were doing to have fun in the park.	3 2 1 0
Beyond the Text Some other things they could do at the park are <u>(examples)</u> . You can do many different things at the park like <u>(examples)</u> . Lots of people like to go to the park to play, have picnics, etc. Note any additional understandings:	Why do people like to come to the park? Can you think of some other things that the boy and his dad could do in the park? Talk about what the boy is thinking about going to the park with his dad.	3 2 1 0

Subtotal Score: 6
 Add 1 point for any additional understandings: 1
 Total Score: 7

Fountas & Pinnell Benchmark Assessment System 1 4

Fun at the Park, Level A, Nonfiction, RW: 24

Student: _____ Date: _____

Draw something you would like to do at the park.

Fountas & Pinnell Benchmark Assessment System 1 5

Fountas & Pinnell Benchmark Assessment System, 1 & 2
Student Summary Sheet

Field Test Coordinator's Name _____ Date _____

Student's ID Number _____


Teacher _____ School _____

Word List
Highest Level Achieved _____ Total # of Correct Words _____

FICTION	
Independent: Level _____	% Accuracy
Comprehension Conversation	0 1 2 3
Fluency	0 1 2 3
Instructional: Level _____	% Accuracy
Comprehension Conversation	0 1 2 3
Fluency	0 1 2 3
Difficult: Level _____	% Accuracy
Comprehension Conversation	0 1 2 3
Fluency	0 1 2 3
NONFICTION	
Independent: Level _____	% Accuracy
Comprehension Conversation	0 1 2 3
Fluency	0 1 2 3
Instructional: Level _____	% Accuracy
Comprehension Conversation	0 1 2 3
Fluency	0 1 2 3
Difficult: Level _____	% Accuracy
Comprehension Conversation	0 1 2 3
Fluency	0 1 2 3

If a child reads more than one book at the independent, instructional or difficult level, then complete an additional summary sheet.

Fountas & Pinnell Benchmark Assessment System 1 4

The logo for Fountas & Pinnell Benchmark Assessment System is centered on a purple rectangular background. The text "Fountas & Pinnell" is written in a white serif font at the top. Below it, the words "A to Z" are written in a yellow sans-serif font inside a yellow oval. At the bottom, "Benchmark Assessment System" is written in a white sans-serif font. A white wavy line is positioned below the text. The purple rectangle is flanked by two horizontal brown bars.

Fountas & Pinnell
A to Z
Benchmark Assessment System

For more information and to review the Fountas and Pinnell Benchmark Assessment System, visit:
www.FountasAndPinnellBenchmarkAssessment.com

The Heinemann logo consists of the word "Heinemann" in a white serif font, centered within a red rectangular background. Below the red rectangle is a blue wavy line.

Heinemann

To Order or more information
www.heinemann.com
Phone: 800.225.5800