The Fountas & Pinnell Benchmark Assessment System is a formative reading assessment comprised of 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 serve as 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.

The Fountas & Pinnell Benchmark Assessment System spans grades kindergarten through 8th grade and are aligned with the A–Z book levels of the F&P Text Level Gradient*, as illustrated in Figure 1. Benchmark System 1 represents levels A–N on the F&P Text Level Gradient*, and Benchmark System 2 represents levels L–Z.

The Benchmark Assessment System books are both “vertically aligned” (they become more difficult as a reader progresses through the levels) and “horizontally aligned” (at each level the fiction and nonfiction books are written at similar levels of difficulty).

FIELD TESTING
Development of the texts for the Fountas & Pinnell Benchmark Assessment System was closely supervised by Drs. Irene Fountas and Gay Su Pinnell, creators of the F&P Text Level Gradient*. 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.

Field testing was conducted with 498 students enrolled in a socioeconomically and ethnically diverse group of 22 schools from five geographic regions across the U.S. Determinations of each school’s socioeconomic status were made using federal guidelines for categorizing low-, middle-, and high-SES schools.

The participating students were from diverse ethnic backgrounds and income groups. Figures 2 and 3 show the student demographics from the field site schools from each state.
Field Study of Reliability and Validity
Benchmark Assessment Systems 1 and 2

EDITORIAL PROGRAM DEVELOPMENT
Drawing upon students’ reading data during the field testing, program developers made changes in the leveled texts to meet the appropriate gradient of difficulty. These changes spanned several dimensions including simplifying the specialized vocabulary words in some nonfiction texts or recasting sentences within a particular text to make them either more or less complex. At one point, it was deemed necessary to replace entirely two texts with more appropriate books. A key change was the establishment of new parameters linking accuracy and comprehension with the independent, instructional, and hard reading levels. This innovative feature provides educators with a more finely grained reflection of a student’s decoding coupled with his or her reading understanding.

RESULTS OF THE TEXTS’ SEQUENTIAL ORDERING
Results from the field testing indicated that the fiction and nonfiction books in the Fountas & Pinnell Benchmark Assessment System progressed in difficulty as the levels increased from Levels A–Z, as depicted in Figure 4 and Figure 5 respectively. By grade level, 84% of the students read the fiction books in a sequential order from lower to higher levels of difficulty within one level above or below the targeted reading level, while 83% of the students read the nonfiction books in that order.

RESULTS OF HORIZONTAL CORRESPONDENCE BETWEEN FICTION AND NONFICTION TEXTS
The field testing also confirmed that students’ developmental reading levels are similar for fiction and nonfiction texts at each level on the F&P Text Level Gradient™. As the chart (in Figure 6) shows, 76% of the students read the fiction and nonfiction books at similar reading levels within one level of text difficulty.

TEST-RETEST RELIABILITY
Test-retest reliability refers to the consistency of students’ scores across tests. 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. In general, test-retest results should exhibit a reliability coefficient of at least .85 for an assessment’s information to be considered stable, consistent, and dependable. As the test-retest results depicted below (in Figure 7) demonstrate, the Fountas & Pinnell Benchmark Assessment System is a reliable reading assessment.

<table>
<thead>
<tr>
<th>Test-Test Reliability Between Fiction and Nonfiction Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book Series A–N</td>
</tr>
<tr>
<td>Book Series L–Z</td>
</tr>
<tr>
<td>All Books (A–Z)</td>
</tr>
</tbody>
</table>

CONVERGENT VALIDITY
The validity of a test is the degree to which an assessment measures what it purports to measure. Convergent validity examines the relationship between an assessment’s scores and scores from other instruments.

There was a strong relationship between the reading accuracy rates of Fountas & Pinnell Benchmark System 1 fiction and nonfiction books (Book Levels A–N), and the accuracy rates of the texts used for assessments in Reading Recovery®, with correlations of .94 for fiction and .93 for nonfiction. This is an important finding because the Reading Recovery® Text Level Assessment, like the Fountas & Pinnell Benchmark Assessment System, assesses decoding, fluency, vocabulary, and comprehension. In addition, Reading Recovery® was recently recognized 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.

There was a moderate association between the Benchmark System 2 (Book Levels L–Z) fiction and nonfiction books and other literacy assessments. One study indicated the Benchmark System fiction texts (correlation of .69) and nonfiction texts (correlation of .62) were moderately related with the Slosdon Word Test. These results indicate that the Benchmark System 2 texts are moderately indicative of the Slosdon measure of word reading. When comparing grade levels, students generally scored higher on the Slosdon than they did with Benchmark Assessment System texts for grades 2–6. However, this pattern was not sustained in grades 7 and 8. The Slosdon Word Test measures students’ isolated oral word calling and provides approximate placement of a child’s reading level. It needs to be emphasized that the Fountas & Pinnell Benchmark System is based on students’ comprehensive reading of complete books.

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 & 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.