

# **STEP Literacy Assessment Technical Report of Validity and Reliability**

David Kerbow and Anthony Bryk

## **Introduction**

Children, particularly in urban settings, begin school with a variety of strengths and challenges that impact their learning. Some arrive with a firm background in early literacy; others have had extremely limited exposure to books. Teachers and school staff have to respond in a strategic way to provide the kind of instructional support that students at these diverse points need in order for them to be successful as readers and writers. This is crucial especially during the early elementary grades as students are developing foundational skills and knowledge that they will build upon as they move across the grades. The diversity in students' beginning points as well as in their trajectories as learners over time places considerable demands on teachers to understand how individual children are processing information, how to organize instruction to accelerate their progress, and how to marshal other resources beyond the classroom to support their growth.

The STEP (Strategic Teaching and Evaluation of Progress) Literacy Assessment was developed by the Center for Urban School Improvement (USI) at the University of Chicago in collaboration with affiliated USI teachers. The assessment provides a set of tools, tightly aligned with scientifically established milestones in reading development, to follow students' progress from kindergarten through third grade. These tools are organized into a developmentally sequenced set of tasks that help teachers understand the developmental status of individual students and a class as a whole at any given point, and to analyze their progress over time. The assessments are woven into classroom practice as an integrated part of literacy instruction rather than a separate activity that is external to teaching.

## Theoretical Background

A number of related and overlapping theories of developmental reading have been proposed in the last 40 years (Bear, 1991; Chall, 1983; Clay, 1991, 2001; Firth, 1985; Fountas and Pinnell, 1996, 2001). Although varying in their details, these theories share in a common view of reading as a complex process in which children learn to combine different sources of information including phonemic awareness, understanding of the alphabetic principle, word recognition and decoding, fluency, and comprehension. Each theory provides description (at various levels of detail) of the skills and strategies that readers demonstrate at each of the developmental stages that they pass through as they learn to read.

In the tradition of these developmental frameworks, the STEP Assessment is organized around a “map” of how students grow as readers. These stages are summarized in Table 1, “Developmental Stages of Reading.”

*Emergent Readers* understand how books work – matching their reading word-by-word in familiar text, often pointing with their finger. They have a growing knowledge of how letters and sounds are associated and can segment the separate sounds in words. In addition, they are putting this knowledge together to help them problem-solve some new words in books by using initial letters and picture support to confirm their reading. Their writing also demonstrates their growing skills and strategies as they correctly spell some high frequency words and represent others with beginning and increasingly ending letters.

*Early Readers* are becoming adept at using letter-sound information along with meaning and language cues to solve new words. They are closely examining letter patterns both in reading and writing as well as developing an increasing number of high frequency and other words that they recognize automatically. Their oral reading, especially with easier texts, is becoming more fluent as they pay attention to punctuation and to expression as they read dialogue. They are also asking questions, making connections with other

information, and working to understand characters' motivation as the plots of the stories they are reading become more complex.

### **Table 1. Developmental Stages of Reading**

#### **Emergent Readers**

*Pre-Reading to Step 2 (Book Levels A-B)*

- Develop the ability to hear separate sounds in words (phonemic awareness)
- Know most letter names and associated sounds
- Read familiar text orally, matching word-by-word (often following with their finger)
- Use information from pictures for understanding and for support in words solving

#### **Early Readers**

*Step 2 to Step 6 (Book Levels B-I)*

- Recognize many high frequency words automatically
- Use letter-sound information along with meaning and language to solve words
- Read easy text with some fluency and attention to punctuation
- Develop the ability to make inferences and interpret text using various strategies

#### **Transitional Readers**

*Step 6 to Step 9 (Book Levels I-M)*

- Use multiple sources of information flexibly while reading for meaning
- Begin to sustain silent reading of longer texts
- Use pictures for information but do not overly rely on them for comprehension
- Build meaning as they read stories, connecting earlier and later parts of a text

#### **Self-extending Readers**

*Step 9 to Step 12 (Book Levels M-P)*

- Problem solve words flexibly with minimal disruption of the flow of reading
- Sustain reading of texts independently over several days
- Try out multiple strategies to support comprehension in difficult text (e.g., test out own understanding by paraphrasing or self-questioning)
- Incorporate new information into their own understanding of a topic while reading nonfiction text

*Transitional Readers* draw from a variety of word-solving strategies and apply them efficiently as they read. They recognize two-syllable words with ease and problem-solve many irregularly spelled words. At these Steps, readers are able to maintain their fluency and reading rate, enabling them to focus more fully on comprehension as they transition into silent reading. Punctuation cues, particularly quotation marks found in dialogue between characters, guide these readers in sustaining meaning. Lengthier texts and more complex plots require Transitional Readers to follow characters through multiple episodes and nonfiction information through multiple chapters. As a consequence, understanding of the story or subject must be built as readers progress through the text. When reading, they combine different details and explore relationships between earlier and later parts of the text, search for pertinent information to support inferences and draw conclusions, and extend their understanding beyond personal experience.

*Self-extending Readers* problem solve words flexibly with minimal disruption of the flow of their reading. Different genres present the reader with new vocabulary and complex words. Context provides some support, as does growing knowledge about word structures, including root words, prefixes, suffixes, and homophones. Self-extending Readers are able to integrate what they read into their own background knowledge and discuss their interpretations and perceptions about stories. They question what authors are trying to convey and why it might be important. In sum, they have developed a system of skills and strategies that propel them to become better readers as they read.

### **Impetus for Developing the STEP Assessment System**

Numerous assessment tasks have been created that directly and indirectly test the separate components that compose the developmental descriptions summarized above. However, these assessments have seldom been combined and sequenced in a comprehensive way nor have the developmental validity of such a system of assessments been examined and validated empirically. Systematic evaluation of the developmental relationships among the components of phonemic awareness, concepts about print, reading fluency, and

comprehension have been lacking. In addition, how such a system of assessments might be adapted for effective classroom use and to help inform teacher practice has been absent as well.

The STEP Assessment system has been purposely designed in response to these concerns. It integrates a set of component assessments around a scientifically established theory of reading development. This assessment system has been subject to extensive analyses and refinements over 10 years of use with teachers across the Chicago region. We summarize below the processes involved in designing and field testing this assessment system, the developmental measurement model that undergirds the system, and the empirical evidence that warrants its use. This system is capable of both tracking typical patterns of student reading growth in early literacy and identifying students whose developmental pattern deviates from the norm, and for whom specialized instructional intervention may be warranted.

### **Comprehensive Literacy Instruction as a Context for Use**

Understanding how readers develop has clear implications for approaches to reading instruction. “Comprehensive literacy approaches” attempt to integrate reading, writing, and word study in explicit response to where students are as readers. In addition to discrete skill development, such instruction aims to help students learn the purposes of literacy and how written language “works.” Students at all levels engage in “learning by doing” through reading and writing in tandem with explicit teaching and guidance from the teacher.

In brief, comprehensive literacy consists of several interrelated elements. Small group reading instruction assures that students learn to comprehend written texts (Person & Fielding, 1991; Pressley, 1998) as well as learn to use phonics skills to take words apart while reading for meaning (Pressley, 1998; Snow, Burns & Griffin, 1998). Instruction is designed to teach comprehension and vocabulary while also providing explicit instruction in reading fluency (NICHD, 2001; Pinnell, et al., 1995). Teachers also provide daily lessons on conventions, skills, and the craft of writing. Students write daily, applying

critical principles to their own production of writing in a range of genres. Instruction in writing, in fact, contributes substantially to children's understanding about words (Clay, 1991; NICHD, 2001) as they learn to hear the sounds in words (phonemic awareness) and learn to look at letters and words (Lieberman, Shankweiler, & Lieberman, 1985; Vellutino & Scanlon, 1987; Lundberg, Frost, & Petersen, 1998) in ways that support both reading and writing achievement.

Teachers marshal the specific activities of the literacy framework in response to students' needs and their understanding of how their particular children are progressing as readers. Authentic opportunities for reading and writing are scaffolded so that students gain increasing independence. Some activities are explicitly demonstrated by the teacher such as shared writing or the interactive read-aloud of books. Others are accomplished with guided support by teachers as students take increasing control of the process as in interactive writing and guided reading. Concurrently, in other activities, children assume primary responsibility through independent writing and reading. These elements come together in the reading of quality children's literature, extensive student writing, and specific attention to letters, words, and how they work (phonics and word study). More than being separate components of a curriculum, the reading and writing activities become a repertoire of practices that teachers weave together based on their pedagogical knowledge and their observation of children.

### **Aims of the STEP Assessment System**

In sum, a comprehensive literacy approach involves complex teaching practices rooted in a deep understanding of the developmental learning process. Rather than being a framework that can be scripted and routinely delivered, teachers must actively develop their classroom practice based on evidence about how their students are actually developing as readers. As a result, reliable and valid assessment of students is not only crucial but central to this teaching endeavor. The STEP Literacy Assessment was specifically designed to support this interactive process of teaching and learning. The design of STEP:

- Builds on a theoretically based description of developmental stages of student reading and empirically establishes a set of diagnostic criteria for understanding their progress;
- Is anchored in a developmental measurement model that allows us to explicitly links critical elements in reading development into an integrated literacy assessment system rather than just providing a collection of loosely related assessment tasks; and
- Provides rich data-based descriptions of students' developmental profiles that facilitates classroom-level instructional planning while also affording more detailed individual student-level data necessary for diagnosing the needs of particular student readers.

## **The Process of Assessment Development**

### **Initial Assessment Components**

The design of the STEP Assessment System takes its root in the Observation Survey developed for Reading Recovery (Clay 1996). These tools are familiar to many teachers who are engaged in comprehensive literacy instruction, although they have not been used systematically in the classroom context itself. The Center for Urban School Improvement first began using the Observation Survey as a research tool for evaluating the effectiveness of its School Development Program that focused on strengthening early literacy instruction. Through this initial work, beginning over 10 years ago, we developed detailed information about the psychometric properties of each of the components included in the Observational Survey. These preliminary studies helped to inform a series of design changes and subsequent field trials that eventually resulted in the current integrated assessment system.

The initial development of STEP drew on several of the separate elements found in the Observation Survey. In the *Letter Identification* task, children are asked to name both

upper and lower case letters. During *Concepts about Print*, students participate in reading a book with the teacher and were asked to indicate directionality from left to right, return sweep at the end of lines, and to point out differences between letters and words on the page. The *Hearing Sounds in Words* task focuses on a student's understanding of letter-sound correspondence and being able to write letters and words from a dictated sentence. The task is scored based on the number individual phonemes the student represents correctly.

The last component that we drew from the Observation Survey was *Leveled Text Reading*. Students read increasingly higher levels of text from pre-primer to fifth grade. Within the Observational Survey, teachers record students' reading accuracy and the assessment ends when a student's reading accuracy for a text level falls below ninety percent.

Each of the separate assessment tasks within the Observation Survey provides valuable pieces of information about emerging student readers. However, it is not necessarily straightforward to analyze how these elements joined to create a complete portrait of a reader that teachers could use to inform instruction. Students exhibit considerable performance variation across these sub-tests even at the same grade level and at the same time of year. A first grade student in the fall of the year might, for example, be near test score floor on one sub-test but close to the test score ceiling on another. In short, the component assessments from the Observational Survey present teachers with considerable data but how best to turn these data into useable information to guide subsequent instruction remained unclear.

Through initial experiences of USI staff in supporting classroom teachers as they sought to use the Observational Survey to inform their own instruction, we gradually came to appreciate the potential utility of a step development framework for organizing the information contained in the various elements that comprise the Observational Survey. Specifically, teachers tend to think about organizing reading instruction for relatively homogenous small groups of students and in this regard the idea of locating students on a

“developmental step” seemed very sensible to them. Moreover, basic research in reading development began to lend some theoretical support for the idea that the discrete skills measured in the Observational Survey might actually align in predictable overall developmental patterns (Stahl and McKenna, 2000; Sainsbury, et al 1999; Ehri, 1995). This prompted us to examine whether the item level data produced by the Observational Study might actually form a developmental scale. Our first attempt at building a developmental measurement model, using a Rasch item response theory methodology provided encouraging empirical results (Kerbow, Gywne, and Jacob, 1999). The items arranged themselves in a sensible developmental order and the psychometric properties of the scale seemed consistent with a theoretical claim that an underlying developmental metric linked these various assessment elements together. We found for example, that patterns of growth in concepts about print aligned with certain levels of ability in hearing sounds in words. These, in turn, were associated with text reading levels of students – all of which could be placed on the same scale. Thus, we began to see the potential (both statistically and practically) for creating an overall system that illuminated the relationship of the different components and the reading skills that they were assessing.

Such a composite scale permits us to move beyond just describing students’ individual scores on isolated assessments to be able to discern where this fit into a larger picture of how students develop over time an integrated ensemble of strategies for learning to read and construct meaning from text. For instructional support purposes, this scaling provides a mechanism for organizing the data from the separate reading assessment tasks into a set of “steps” that offer both a concise summary of students development to date and an interpretive context that could meaningfully inform subsequent instructional decision making. For research purposes, this composite scale affords us with a single continuous measure for assessing students’ overall reading improvement across the primary grades. This solved the typical “floor and ceiling effects” problems typically associated with analyzing data from the separate assessment tasks within the Observational Survey.

In addition, as the empirical basis for the developmental framework began to emerge, our initial analyses also pointed to areas where the assessment information was thin or, in

some case, absent on key concepts. Thus, our analyses of the relative strengths and weakness in the Observational Survey, as discussed below, lead to the design of the final STEP Literacy Assessment System described in this document.

### **Expanding the Assessment Components**

*Broadening the Data on Phonological Awareness and Phonics.* The Letter Identification task from the Observational Survey provided reliable and essential information of students' early development as readers. In our preliminary studies we found that as children's knowledge of letters increased, their understanding of books as measured by the Concepts about Print assessment also developed. However, for some children their knowledge of letter names was not necessarily accompanied by the ability to articulate a letter's associated sound. This distinction represented an important gap in our initial assessment protocol.

The Hearing Sounds in Words task did provide some important information about letter-sound association and phonemic awareness. Students were representing the sounds that they heard in words with letters or groups of letters in their writing. They also wrote some high frequency words in the dictated sentence from memory.

When we began to look at these aspects developmentally, however, several key elements needed further attention. We had limited information, for example, about how students were able to orally identify and manipulate sounds such as hearing onset and rime or separating the first sound they heard in words from the remainder of the word. In addition, as children's understanding of sound and letter association grew more sophisticated, it became clear that how *patterns of letters* related to sounds in words needed to be followed more systematically. Children were asked to write several of these patterns in the Hearing Sounds in Words assessment but the patterns did not build systematically in the sentences used in the Observation Survey.

Thus, though the Observation Survey provided some basis for following a student's development in phonemic awareness and phonics, it became clear that additional elements had to be added in this domain.

*Supplementing Reading Accuracy with Data on Fluency and Comprehension.* The core of our assessment was built initially around the idea of level texts, organized similar to an informal reading inventory. The level texts were initially drawn from the text that Reading Recovery used in connection with the Observation Survey. However, the Observation Survey only assesses reading accuracy on leveled text. Early in the process the need for more explicit focus on comprehension became evident. Specifically, we found that the text level reading accuracy was a strong predictor of comprehension through the early levels (approximately through early second grade). However, at higher levels, students with similar reading accurately showed a much wider variance in their comprehension scores on standardized tests. Thus, although reading accuracy appeared to be a necessary component for early comprehension, it was not sufficient in predicting later stages of reading development. As the text levels became more difficult, reading accuracy became less predictive of a student's understanding of the reading material. Students could decode the words, but not necessarily make meaning from the text.

In response, we incorporated several additional components. An essential link between reading accuracy and comprehension is rate and fluency of the student's reading. Students who spend an inordinate amount of mental energy decoding words tend to lose the broader meaning of sentence, paragraphs and the larger text. Excessive attention to word level concerns leaves limited space for memory and other cognitive process involved in making meaning. Thus, adding a direct recording of reading rate and a fluency rubric (i.e., a four-point scale of students' phrasing and expression during oral reading) became crucial to understanding the reading process at this developmental stage (Pinnell, et al., 1995).

Similarly, we added a set of oral questions linked specifically to each leveled text used in the assessment system. A student now reads the book aloud for purposes of assessing

accuracy, rate and fluency and then, upon completion of the read aloud, engages in a structure comprehension conversation with the adult who is conducting the assessment. These comprehension questions probe both the student's explicit and implicit understanding of the text.

Thus, in addition to reading accuracy, the STEP assessment system also evaluates fluency and comprehension within an overall developmental framework of the reading process.

### **Summary of Results from Preliminary Reliability and Validity Studies**

Full details about the preliminary studies that helped to inform the final design of STEP can be found in Appendix A. In brief, we learned the following from these pilot research and development activities:

- It is possible to combine data from the Letter Identification task, Concepts about Print, Hearing Sounds in Words, and Text Reading Accuracy into an internally coherent developmental scale with high overall reliability (Cronbach's alpha of 0.94). Most important, the scaling of the relative difficulty of the items followed closely with that predicted by development reading theory. This was our first empirical evidence supporting the construct validity of an overall developmental scale. The subsequent addition of items assessing fluency and comprehension also scaled in ways consistent with the idea of a single, underlying development metric.
- The STEP developmental scale scores also displayed concurrent validity when compared against a standardized reading assessment measure, the Degrees of Reading Power (Koslin, Zeno, Koslin, 1987; DRP Handbook, 1995). In spring of first grade, a classroom administered STEP correlated 0.51 with DRP reading scale. For spring second graders, the correlation of STEP with DRP was 0.62. These moderately strong correlations were encouraging. While STEP and DRP tap related content, STEP includes a more expansive array of developmental tasks not typically included in standardized reading assessments. Under these circumstances, moderately strong correlations are exactly what we would expect.

- The developmental scaling also provided an empirical basis for establishing a set of cut scores that define 13 different step levels. Each step level represents a distinct profile of student performance across the multiple reading tasks that comprise the assessment. The development of these step levels subsequently proved highly useful to teachers in planning instruction, and in make decisions about student referral and placement for supplemental reading services.
- Finally, the pilot developmental assessment helped to establish step criterion-based benchmarks for kindergarten, first, and second grade. By comparing students step level against the likelihood of scoring at or above grade level on a spring administered DRP we were able to establish STEP performance benchmarks for kindergarten, first and second grade. For example, in second grade only 36 percent of the student scoring at step 8 in May scored at or above grade level on a concurrently administered DRP. In contrast, 79 percent of the students at step 9 scored at or above grade level on the DRP. With this kind of empirical evidence, we established step 9 as the “developmental benchmark” for end of second grade. Similar analyses led us to establish step 6 as the benchmark for the end of first grade.

## **An Overview of the Current Assessment**

### **Overall Organization of the Assessment**

The assessment system is organized around a developmental trajectory consisting of 13 distinct steps from pre-literacy to Step 12 that provide a comprehensive map of students’ reading skill acquisition. This generally maps on to the skill development expected of students during the period from K to end of grade 3.<sup>1</sup>

---

<sup>1</sup> STEP contains two approaches that provide a window into evaluating how students are progressing. The first, described here, draws on a set of *formal evaluations* that are individually administered and organized around a set of leveled books. The second uses *informal observation checklists* that create a lens to look at student reading behaviors during classroom and tutoring activities. Combining the formal and informal assessment provides a complete portrait of students’ strengths and weaknesses to support instruction, grounded in the same understanding of developmental reading. The informal observation checklists are still under evaluation.

Central to the assessment is a set of leveled texts that increase in difficulty with each “step.” During individual conferences of 10 to 15 minutes, the teacher records students’ reading accuracy and fluency, observes their reading behaviors, and engages students in comprehension conversations about what they have read. Importantly, however, each Step, in conjunction with the leveled books, also includes assessment tasks that provide a deeper look into specific skills that supplement what is learned from students’ oral reading. That is, STEP explicitly joins the reading of authentic texts with assessments that focus on level appropriate individual skills such as letter-sound association, phonological awareness, and word knowledge -- providing a complete window into the integrated development of the reading process.

Thus, at the heart of STEP is a set of empirically-grounded relationships about the specific strategies and skills students acquire as they read and understand increasingly complex text. A demonstration of the ability to read and understand text at each level, in essence, represents “steps” in students’ development as readers. The formal assessment provides systematic information about the integrated development of this process by focusing on key components that research shows are essential to assessing the building blocks to reading:

- Concepts about print
- Letter name and sound knowledge
- Phonological awareness
- Reading accuracy
- Reading rate and fluency
- Comprehension
- Developmental spelling

### **Organization of the Assessment Components by Step**

Table 1, “Components of the Assessment,” delineates how each component or task in the assessment system is organized by step level. Many of these tasks are similar to other early literacy assessments contained in DIBELS, the Early Reading Screening Inventory

(ERSI), the Developmental Reading Assessment (DRA), and the Texas Primary Reading Inventory (TPRI). However, the explicit combination of these tasks in developmental sequence is unique to the STEP Assessment and is organized on both a theoretical as well as an empirical basis. (The underlying measurement model that confirms the validity and reliability of this organization is discussed in the next section.)

**Table 1**  
**STEP Assessment Components**

	Pre-Reading	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10	Step 11	Step 12
<b>Story Retelling</b>									✓	✓	✓	✓	✓
<b>Comprehension Questions</b>				✓	✓	✓	✓	✓	✓	✓-3 written	✓-3 written	✓-3 written	✓-3 written
<b>Reading Rate and Fluency</b>				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<b>Text Level</b>		✓-A	✓-B	✓-C	✓-E	✓-G	✓-I	✓-K	✓-L	✓-M	✓-N	✓-O	✓-P
<b>Developmental Spelling</b>		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
<b>Phonemic Awareness</b>													
Segmentation			✓	✓									
First sounds		✓											
<b>Letter Identification</b>													
Sounds		✓	✓	✓									
Names	✓	✓	✓										
<b>Concepts about print</b>	✓	✓											
<b>Phonemic Awareness</b>													
Rhyming Words	✓												
<b>Name Assessment</b>	✓												

*Alphabet Recognition.* The single best predictor, on its own, of early reading achievement is accurate naming of the letters of the alphabet (Adams, 1990; Snow, Burns, and Griffin, 1998). The Name Assessment task provides the first indication of students' knowledge of letters and letter names that will be building more generally across Step 1 and Step 2. The assessment asks the student to write his or her first and last name and, then, to identify several letters in this very familiar context.

The *Letter Name Identification* task asks students to provide the names of both upper and lower case letters.<sup>2</sup> Students are presented with a page of letters (upper case then lower case) that are in random order. They are asked to point and say the name of the letter. This is not a timed task although if students take more than five seconds with a letter they are prompted to move to the next letter. The Letter Name Identification assessment extends from the Pre-reading Step to Step 2.

Pronouncing the sounds represented by individual letters in isolation is a more demanding task for students. It requires explicit awareness of individual phonemes and their association with letters. The *Letter Sound Identification* task uses the lower case letters, asking students to touch each letter and say the sound it represents. Teachers ask for the alternate sound for a letter that has two (or more) sounds. The focus for vowels is the short sound, and the hard sound for 'c' and 'g.' For example, a child that gives the long vowel sound would be prompted: Do you know the other sound the letter can make? Do you know the short sound this letter makes? This assessment extends from Step 1 to Step 3.

Taken together these assessments provide a complete sense of a student's alphabet recognition beginning in the familiar context of his or her name and proceeding to recalling letter names in isolation and producing their associated sounds.

---

<sup>2</sup> The lower case letters include both a print and typescript version of 'a' and 'g' making a total of 28 responses.

*Concepts about Print.* Understanding how books work is a foundational knowledge that supports children’s access to text. The emergent reader begins to recognize directionality of the text from left to right, at the end of a line to return to the beginning of the next line and then on to the next page, and to start “reading” at the top left of the text. These early concepts about print provide a context for using their building knowledge of letters and sounds.

Developmentally, these early concepts of how text works progress to an understanding of the concept of a word. Ultimately, this refers to the emergent readers’ ability to match spoken words to written words as he or she reads. This involves the integration of alphabet recognition, letter sounds, and word boundaries in text. Research has shown that a stable concept of word in text facilitates a child’s awareness of the individual sounds within words. Until a child can point to individual words accurately within a line of text, he or she will be unable to learn new words while reading or to attend effectively to letter-sound cues at the beginning of words (Morris, 1993; Morris et al., 2002).

Embedded in the *Concepts about Print* assessment is a series of probes that allow the student to demonstrate this growing understanding of the concept of a word as well as one-to-one matching of words to print.

*Phonological Awareness.* Phonological and, more specifically, phonemic awareness plays a key role in literacy development (Adams, 1990; Whitehurst and Lonigan, 1998). Children who are better at detecting rhymes or phonemes are quicker to learn to read – even after other factors such as vocabulary, memory, and socioeconomic status are taken into account (Wagner and Torgesen, 1987).

Recent research suggests that phonological awareness can be best conceptualized as a single underlying ability that increases in complexity as readers develop (Anthony and Lonigan, 2004). STEP employs three assessment tasks that span these levels of difficulty in order to follow students’ development in this essential skill.

Phonological/phonemic awareness is assessed across several task of increasing complexity. Components include: Rhyming (Pre-Reading), Matching First Sounds (Step 1), and Segmentation (Step 2-3).

*Developmental Spelling.* Application of letter-sound knowledge in invented spelling tasks is an excellent predictor of word recognition in young children (McBride-Chang, 1998) and among the best predictors of word analysis and word synthesis (Torgesen and Davis, 1996).

Research on how children learn to read and spell words in an alphabetic orthography has consistently revealed that orthographic features are internalized for reading and writing in a systematic, developmental progression. The process of learning to read and learning to spell are intricately linked (Ehri, 1997). Therefore, assessing spelling provides information about students' understanding of orthographic patterns not only in writing words but also in reading. Invented spelling provides a diagnostic window into students' understanding of alphabetic orthography and can help teachers determine when to teach what phonics or spelling feature of English orthography (Henderson, 1990). According to this body of research, the acquisition of basic phonics features occurs in a predictable progression: beginning consonants; ending consonants; consonant diagraphs, medial short vowels; consonant blends and pre-consonantal nasals; silent-e marker for long vowels; other long vowel patterns; r-controlled vowel patterns; then ambiguous vowel-diphthongs and digraphs (Bear, et al 2000; Ganske 2000). Table 2 shows how these features are organized by step within the overall assessment system.

Words were selected to represent each feature based on two criteria. First, the word should typically be in the oral vocabulary of children at each level. Second, the word should not be of such high frequency that it would likely be in most children's writing vocabulary with automaticity. The later criterion ensures that students' must attend to the features of the word during the assessment. The assessment is scored based on the student's ability to correctly spell the specific feature being tested. This in turn provides

teachers with specific information about the spelling patterns that students have under control at each step.

**Table 2. Developmental Spelling across Steps**

<p><b><i>Steps 1-3:</i></b> Early to Middle Letter Name-Alphabetic Stage</p> <p>Focus: Beginning, ending consonants; early short vowel recognition</p> <ul style="list-style-type: none"> <li>• Most beginning and ending consonants</li> <li>• Clear letter sound correspondence</li> <li>• Representing short vowels in C-V-C words (e.g., may write ‘pat’ for ‘pet’)</li> </ul>
<p><b><i>Steps 4-5:</i></b> Late Letter Name-Alphabetic Stage</p> <p>Focus: consonant digraphs; short vowels</p> <ul style="list-style-type: none"> <li>• Regular short vowel patterns</li> <li>• Most consonant blends and digraphs (e.g, frog, ship)</li> <li>• Long vowels often represented by letter name (e.g., may write ‘tran’ for ‘train’)</li> </ul>
<p><b><i>Steps 6-7:</i></b> Early Within Word Pattern</p> <p>Focus: beginning long vowel patterns; r-controlled vowels</p> <ul style="list-style-type: none"> <li>• Using some long vowel patterns, especially V-C-e</li> <li>• Beginning accuracy on r-controlled vowels in single-syllable words (e.g., bird, fur)</li> <li>• Consistent with short vowels, blends, and digraphs</li> </ul>
<p><b><i>Steps 8-10:</i></b> Middle/Late Within Word Pattern</p> <p>Focus: long vowel patterns; r-controlled vowels; vowel digraphs</p> <ul style="list-style-type: none"> <li>• All of the above plus:</li> <li>• Single-syllable long vowel and r-controlled words</li> <li>• Many common vowel diphthongs (e.g., round, point)</li> </ul>
<p><b><i>Steps 11-12:</i></b> Early Syllables and Affixes</p> <p>Focus: -ed/-ing endings; consonant doubling; vowel patterns in syllables</p> <ul style="list-style-type: none"> <li>• -ed and most inflections</li> <li>• Some vowel patterns in two to three syllable words</li> <li>• Consonant doubling (e.g., shopping, tennis)</li> </ul>

*Text Level Reading.* Listening to student read aloud from leveled text provides direct information for understanding their reading skill and strategies, diagnosing strengths and weaknesses, and evaluating progress (Johnson, Kress, and Pikulski, 1987). Developing a set of leveled books with appropriate layout and content is essential to this process. As teachers become familiar with these books, they are able to identify key aspects of the texts that reveal students' approach to problem-solving words and making meaning.

Multiple approaches have been developed in attempting to determine the reading difficulty of books (Hoffman, et al, 2000). Readability formulas have proliferated and can provide a general guide. However, they can be widely variable in how they rank text, particularly at early levels. Books that support developing literacy (and thus are most appropriate for assessment) contain several crucial characteristics (Fountas and Pinnell, 1996). These include:

- A mixture of natural and literary language patterns;
- An increasing number of high frequency words;
- Opportunities to notice and use spelling patterns within words;
- An increasing syntactical and grammatical complexity in sentence; and
- Interesting and engaging content for children.

Building on the text reading gradient developed in Reading Recovery, Fountas and Pinnell (1996, 1999) propose a leveling system that incorporates the above characteristics to define "benchmark" books. A benchmark book is a reliable exemplar for a particular level in the gradient of difficulty. It has been judged to be readable, at or above 90 percent accuracy, by most students who demonstrate similar reading behaviors at a particular point in time. For example, a benchmark book at level C is read accurately by children who are using first sounds, recognizing several high frequency words, and build meaning using pictures and natural language patterns. Thus, there is an iterative process in defining a books level: examining the characteristics of a book, observing children read the book who are developmentally similar readers, and comparing the book to other

books those children can read. From this process, they have described a series of levels using a letter system ranging from A to P.

It is this process and leveling system (which is closely tied to comprehensive literacy instruction) that was employed in developing the leveled text for the STEP Assessment. The validation of the leveling includes computation of Lexile readability scores to establish general ranges of reading difficulty, comparison to Reading Recovery text, and field testing of the books with students at various achievement levels. The results are detailed in the next section.

*Accuracy, Reading Rate, and Fluency.* Accuracy and speed may be considered essential building blocks of reading fluency. If readers have too much difficulty recognizing and reading individual words, their ability to gain overall meaning from a passage will be seriously hampered (Samuels 1994). As readers pause to problem-solve unfamiliar words, the ideas that were developing within the sentence or across the portion of the passage may be disrupted – having a negative impact on students’ ability to understand the text. If reading proceeds too slowly or mechanically, these connections may become difficult or impossible to make.

The essential building block of accuracy can be measured in a straightforward manner through counting the number of errors (i.e., words read incorrectly or skipped) and comparing this to the number of words read in the text. However, even competent readers may make periodic deviations from the text. In some cases, these deviations may signal a reader’s attempt to make sense of what is being read. In other instances, the errors may be an indication of a further break down of the reading process. Careful analysis of oral reading, thus, includes both attention to the accuracy of the reading but also the type of errors and substitutions that readers are making (Clay, 1996; Hennings & Hughes 1992).

Reading rate is also directly measurable through timing oral reading and computing the number of words read per minute. Traditionally, these two aspects of oral reading – accuracy and rate – have been the primary indicators of oral reading ability. They

logically form the foundation for the ability to read with fluency, that is, with phrasing, attention to syntax, and appropriate expression associated with the text.

To develop a simple rating scale for fluency that incorporated aspects of pacing, phrasing, and expression, we adapted the four-point fluency rating scale used by the National Assessment of Educational Progress (Pinnell, et al., 1995). The rubric is displayed in Table 3.

In summary, the STEP Assessment pays close attention to the key aspects of strong oral reading -- accuracy, rate, and fluency – to present a well-rounded view of this central element of the reading process.

**Table 3. Fluency Rubric**

<b>Level 4</b> Reading in meaningful phrases; consistently pays attention to punctuation and syntax; reading some or most of text with expression; may slow briefly for problem-solving, but quickly returns to fluent reading
<b>Level 3</b> Reads primarily in 3-4 word phrases; pays attention to punctuation and syntax most of the time, but read with little expression; may occasionally slow for problem-solving
<b>Level 2</b> Reads primarily in 2-3 word phrases; seldom pays attention to punctuation ad syntax; slow problem-solving fairly often
<b>Level 1</b> Read primarily word-by-word (may reading fast or slow, but rhythm is word-by-word); slow problem-solving

*Comprehension.* Cognitively based views of reading comprehension emphasize the interactive nature of reading (Rumelhart & Ortony, 1977; Dole, et al., 1991) and the constructive nature of comprehension (Pressley 1998). In order for assessment to explore this process in young readers, teacher have to engage students in conversation about how their understanding of a text as well as how they support their conclusions. This

“comprehension conversation” is organized around a set of questions in the assessment that span a range of cognitive demands as described in Table 4.

The teacher’s role in this process is not simple one of passive questioner and listener. Instead, she plays the role of discussant with the student, using a structured set of questions as well as non-leading prompts to encourage students to further explicate their thinking and basis for understanding.

Appropriate prompts do not lead children to an answer, but rather encourage them to elaborate or explain their thinking. To elicit student’s responses, the following prompts help to review students’ prior knowledge, how they are making sense from the text, and what they consider to be important evidence.

- What in the book makes you think that?
- Tell me more about . . .
- What do you mean when you say . . .?
- Why do you think that?
- Why do you think it’s important that . . . ?

This approach inherently requires that questions be open-ended. Consequently, teachers have to evaluate whether responses show clear understanding. Although sample answers are provided, judgment about appropriate probes as well as quality of response is still necessary. For each group of teachers, this necessitates conversation among themselves to establish reliability in administration and scoring. Rather than becoming an impediment, the social practice of establishing reliability of scoring pushes teacher to explore what they perceive to be evidence of student comprehension as well as what are effective questioning and prompting techniques during classroom instruction.

The comprehension questions begin with the early level books at Step 2. As the complexity of the books increase across steps more elaborate response are possible. At Step 9, in addition to the oral questions, students respond in writing to three questions.

This is an important skill assessed on many standardized tests and allows the teacher to evaluate students' ability to organize their ideas in writing as well as to elaborate without prompts.

**Table 4. Question Categories**

<p><b><i>Factual/Literal Questions</i></b> have only one correct answer and are used to assess whether students understand the literal meaning of a text. Factual questions might ask students to:</p> <ul style="list-style-type: none"><li>Recall details;</li><li>Recap descriptions of people, places, or events;</li><li>Explain relationships among characters; and/or</li><li>Establish time and sequence.</li></ul>
<p><b><i>Inferential Questions</i></b> generally have only one correct answer. They require students to look at the text for clues that will lead to the answer and are used to assess whether students can use the text to:</p> <ul style="list-style-type: none"><li>Draw logical conclusions about literal meaning;</li><li>Deduce the meaning of words through context;</li><li>Understand cause and effect;</li><li>Understand figurative and metaphorical language; and/or</li><li>Identify problems and solutions.</li></ul>
<p><b><i>Critical Thinking Questions</i></b> can elicit a variety of answers. They require students to draw upon their powers of analytical and imaginative thinking and use the text to:</p> <ul style="list-style-type: none"><li>Explore characters' traits, feelings, and motivation;</li><li>Understand the story's mood or tone;</li><li>Articulate central ideas or themes in a story;</li><li>Originate and support ideas;</li><li>Use textual evidence;</li><li>Draw conclusions about an author's viewpoint, or intention; and/or</li><li>Entertain differing interpretations.</li></ul>
<p><b><i>Personal Opinion Questions</i></b> ask students to move beyond analysis of the text. They require students to apply their own values and experiences in evaluating their feelings about what they have read and:</p> <ul style="list-style-type: none"><li>Consider if they agree with what characters in the story think or do;</li><li>Empathize with or judge characters; and/or</li><li>Decide whether they agree or disagree with the author.</li></ul>

## **Description of Psychometric Studies**

As noted earlier, the STEP assessment system emerged out of a design research effort that involving an interplay of co-design work with classroom teachers and literacy coaches using pilot instruments and interpreting results with ongoing small scale studies of the technical properties of the evolving instruments. As the design for the assessment system took final form, we undertook two formal studies to evaluate the reliability and validity of STEP when used in ordinary classroom settings to inform future instruction and supplemental programming for students. The design for these two studies is summarized below.

### **Construct Validity-Internal Reliability Study**

This study served several purposes. First and most importantly it provided data to evaluate the validity of the overall developmental scale. Second, it provided information about the internal reliability of the overall scale and the reliability of its sub-components(e.g., letter name identification, segmentation, etc) in *regular classroom use*. Data were collected in classrooms by regular teachers as part of their normal instructional activities with their own students. Although all teachers were trained to use the assessments, no other standardized procedures were imposed on test administration. Third, since two different forms of STEP currently exist (a “purple” and a “yellow” form), this study was also designed to examine the parallelism of these two alternative versions.

The data for this study were collect during the spring of the 2003 academic year by classroom teachers in six Chicago schools with predominately low-income and minority student populations. Table 5 displays the sample size for this study by grade level.

**Table 5. Construct Validity-Reliability Dataset**

Grade	Frequency	Percent
K	60	14%
1	110	26%
2	121	29%
3	130	31%
Total	421	

The students in this study were 81% African American and 19% Latino. Some 85% were from low income family eligible for free and/or reduced price school lunch. These schools, and their student population are modal within the Chicago Public School district.

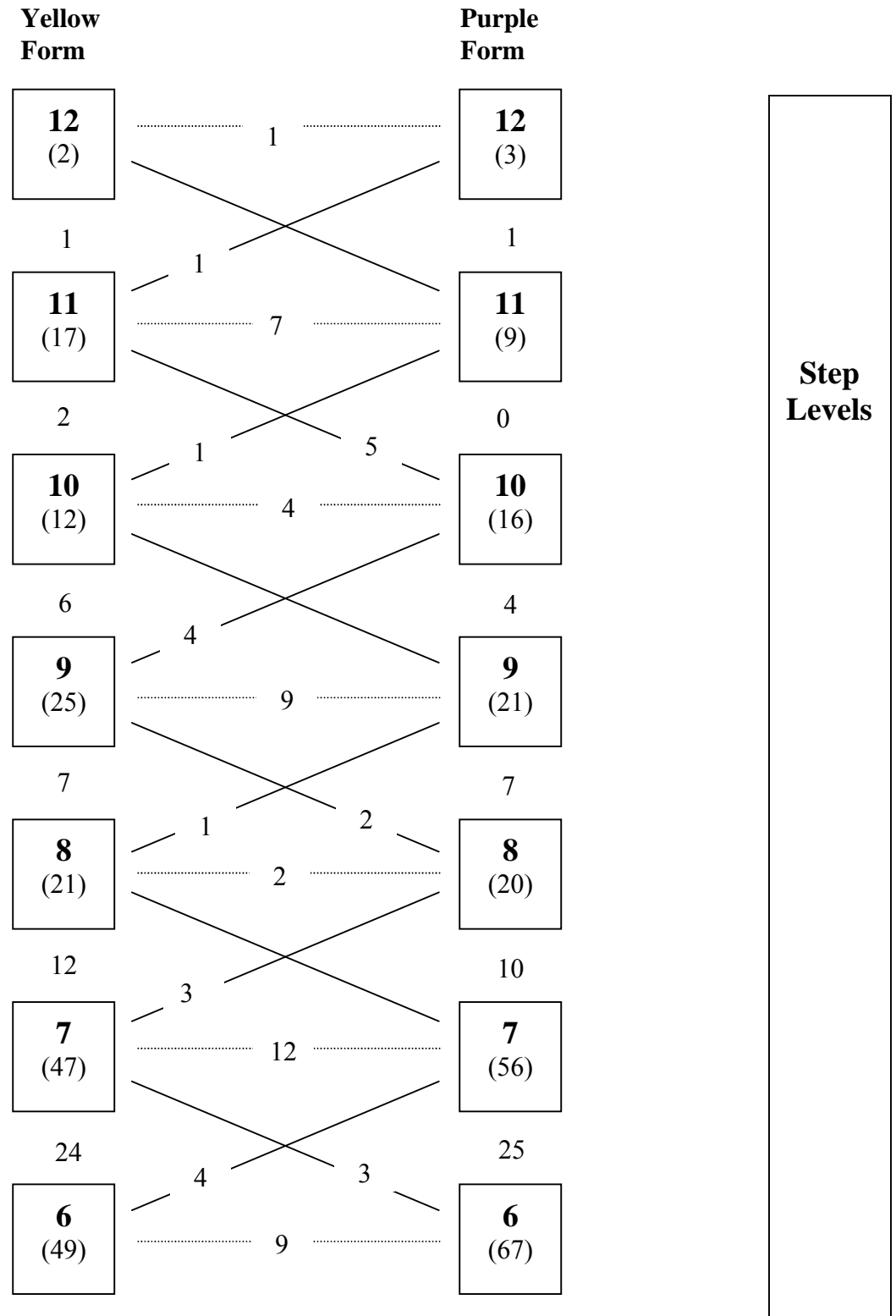
The data consists of item response level information for each assessment task administered to this student sample. Based on students' prior STEP results from earlier in the academic year, teachers were asked to administer the next appropriate step level to each child. Since one primary aim of this study was to evaluate the developmental linkages across the 13 steps that comprise the assessment system, it was necessary for some students to be assessed across two or more steps. These students took both their "next step level test" plus the one immediately above it. In addition, in order to evaluate the parallelism of the two forms, it was also necessary for some students to take the same step level on both the "yellow" and "purple" version of the test. For the students who took these double assessments, the first and second administrations were separated by a one week period of time.

Figure 1 displays how students were distributed in this study across the various step levels and forms. The numbers in parentheses within the boxes are students who took only one step at the time of administration. The numbers that link boxes are students who took two step assessments. For example, 47 students took step 7 of the yellow form; 24 students took both steps 6 and 7 of that form, and 12 students took step 7 from both

forms. Another 4 students took the combination of step 6, yellow and step 7, purple, and 3 students took the combination of step 6 purple and step 7 yellow.

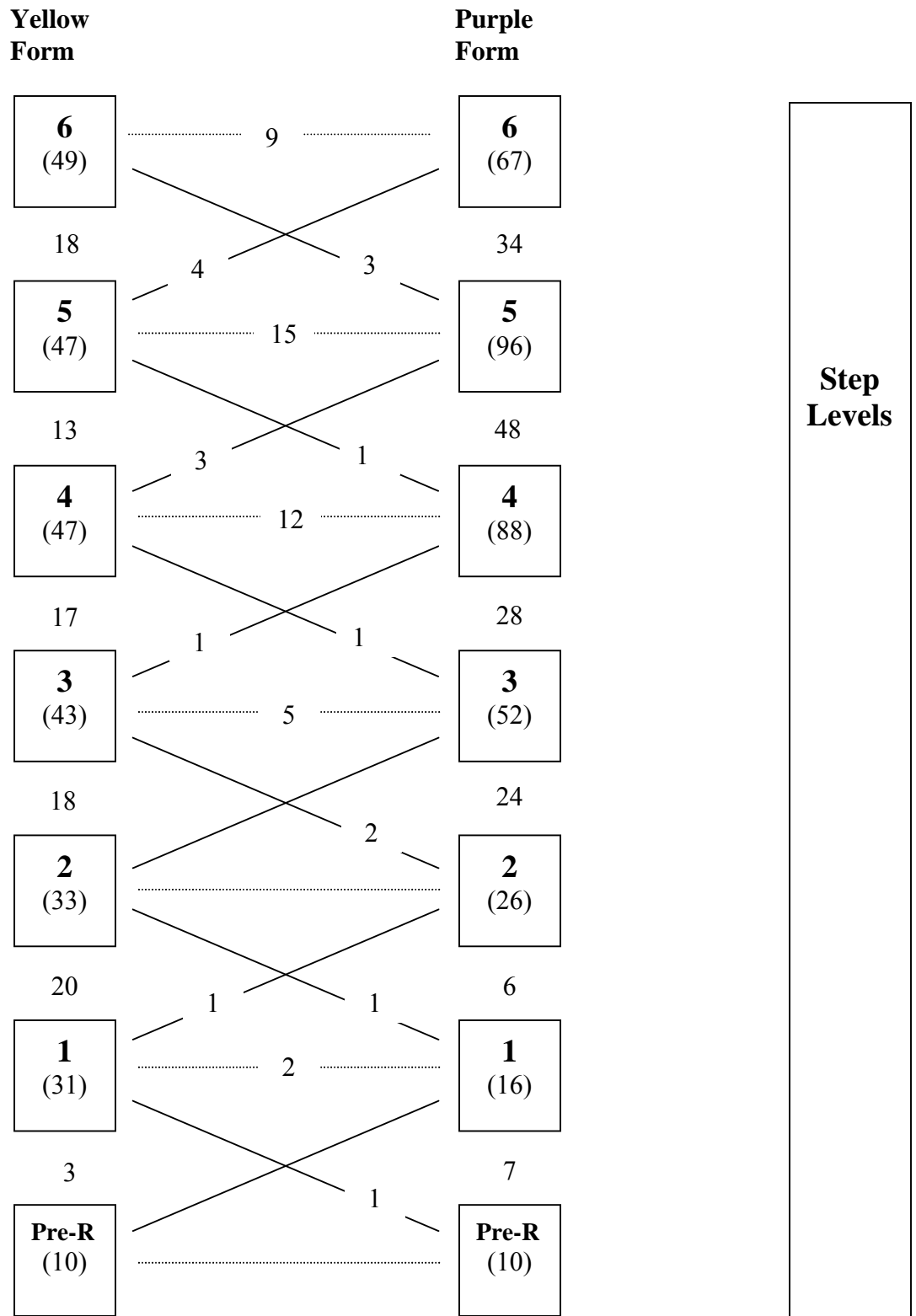
In general, the final data set has a relatively high density of observations for steps 1 through 11, but is thin at steps “pre-reading” and 12. Since a key objective of the study was to assess reliability and validity in the context of classroom use, the data collection was organized around in-tact classrooms. By the spring of kindergarten, relatively few students remained at the pre-reading level. In contrast, relatively few third graders were ready for step 12 at the time of the study since this was a very disadvantaged urban school population. Fortunately, since there is significant item overlap across the levels and forms of the assessment, this data thinness at the lower extreme does not material effect any of the results presented below. Interpretation at Step 12 requires more caution.

**Figure 1. Equating Study Design: Linking Students across Steps and Between Forms**



Note: The (xx) inside each box indicate the number of students who took that form and step of the test. The counts between the boxes indicate the number of students who took that particular combination of steps and/or forms of the assessments in the construct validity/internal reliability study.

Figure 1 – continued



## **Concurrent-Predictive Validity Study**

This study focused attention on examining the concurrent and predictive validity of the individual step designations which result as an end product of a STEP assessment administration. For a student to be designated as “achieving step level”, he or she must meet a performance target on each assessment component that comprises that step. These target levels were established during the co-design process with classroom teachers and literacy coordinators. (For further details see Kerbow 1999. Evidence on the empirical validity of these step targets is presented later in this report.)

Our concurrent-predictive validity study examined the alignment between these step designations and students concurrent and subsequent performance on two different standardized reading tests. Specifically, we compared students end of year STEP level results with their reading achievement scores from a spring administration of the Iowa Test of Basic Skills (ITBS) in second and third grade and with student results from the required state accountability reading assessment at grade 3, the Illinois Standards Achievement Test (ISAT) In each case, the STEP assessments were administered by regular classroom teachers and results obtained prior to the standardized tests.

This dataset consists of 4 cohorts of students from a single Chicago public school serving 100 percent African American students who were 75 percent low income. Each cohort was followed for a full year from spring of second grade through spring of grade three. Both the STEP Assessment designations at grades 2 and 3 and results from the two standardized reading assessment were recorded. Figure 2 illustrates the overall design for this study. .

**Figure 2**  
**Concurrent-Predictive Validity Study**

	<b>Spring 2000</b>	<b>Spring 2001</b>	<b>Spring 2002</b>	<b>Spring 2003</b>	<b>Spring 2004</b>
<b>2<sup>nd</sup> Grade</b>	24	25	25	47	48
<b>3<sup>rd</sup> Grade</b>		22 (2)	24 (1)	18 (7)	40 (9)

The first row in the figure shows the number of second grade students assessed each spring. The second row displays the number of students who continued with their cohort into grade 3 the following year (e.g., 22 of the original 24 students from the 2000 cohort were retested in spring 2001). In parentheses are the number of students who joined the cohort for the first time in third grade (e.g., 2 in spring 2001). The total dataset consists of 188 students.

### **Construct Validity of STEP as a Developmental Assessment**

As noted in the introduction, STEP was intentionally designed to integrate a set of component assessment tasks, each of which has a separately established scientific basis. The design for the STEP assessment system has been explicitly organized around an overall theory of reading development rooted in several decades of scientific research. Thus, there are good theoretical reasons for hypothesizing that the various assessment

tasks we have assemble would integrate into a single developmental scale. The Item Response Theory analyses described below provide key empirical evidence about the adequacy of STEP as an overall development measure rather than as a loose collection of discrete reading tasks. While both theory and previous empirical findings in the literature guided our assessment development work, whether STEP actually functions as a single integrated scale is ultimately a question about its internal psychometric properties.

### **Use of an Item Response Theory Model to Examine Developmental Scale Properties**

In order to investigate the adequacy of STEP as a development scale, we applied a Rasch analysis to the equating design data previously detailed in Figure 1. The Rasch model (Wright & Stone, 1979; Wright & Master, 1982) is a simple form an of item response latent trait model which is now commonly employed in standardized test construction. Test items are used to define a measurement scale based on the relative probability of a respondent's correct answer to each item. In a properly fitting Rasch scale, the item hierarchically arrange in difficulty order, and individuals are "measured" on this same scale based on their responses to particular sub-set of items administered to them. Individuals do not need to take all of the items that comprise the scale. A valid measure (i.e. scale score) can be assigned based on a child's responses to the subset of items actually administered. The scale units are measured in logits, (i.e. the log odds of a correct response), and provided the data meet the assumptions of Rasch theory, the scale constitute a linear measurement system suitable for use in standard statistical procedures.

A Rasch analysis produces a diverse array of statistics for examining the quality of the underlying measurement system. First are item difficulty statistics, which estimate the likelihood that respondents will produce a correct response for each of the items that compose the overall scale. For example, easier items such as letter identification should tend to have lower difficulty estimates than more advanced items such as comprehension questions associated with higher level texts. Thus, how the item difficulties locate themselves within the overall scale provides critical evidence as to whether the

component tasks, which comprise the integrated scale, are empirically sequencing in the expected theoretical order.

Second, and also important are a set of item infit statistics. These estimates provide information about the degree to which individuals' responses to a particular item are consistent with its placement in a hierarchically ordered scale. If STEP is truly a developmental scale, individuals who answer correctly a particular item should be more likely to answer correctly the easier items below it in the scale, and be less likely to answer correctly the harder or "more difficult" above it in the scale. In essence, the infit statistic measures the degree to which the item's behavior deviates from what we would expect in a perfect developmental scale. Under the Rasch model, a probability value can be attached to the estimated infit statistic associated with each item. The underlying null hypothesis associated with this test statistic is of the form "assuming that this item is part of a hierarchically ordered scale, how likely is it that we could get an infit statistic this large by chance alone?" By tallying the percent of items that exceed an expected  $\alpha$  value of .05, we derive empirical evidence for examining whether the tasks and sub-components cohere as an overall developmental scale. More specifically, within any cluster of items we expect some individual items to misfit to some degree as a matter of statistical error. Evidence of systematic misfit however, e.g. a relatively high incidence of misfits among the items in the spelling assessment task, would indicate that this component follows a different development pattern than the other items in the scale. Thus, this aspect of the Rasch analysis provides critical information for judging whether it is appropriate to view the combination of sub-components as forming one underlying scale, or alternatively whether some more complex multi-dimensional developmental scaling is needed.

Third, the Rasch analysis provides a standard set of statistics about the internal consistency of the overall scale, and of the specific sub-components that form this scale. The "person reliability statistic" generated in a Rasch analysis is an internal consistency reliability measure similar to Cronbach's alpha coefficient.

Fourth, although not of primary concern for the psychometric studies reported here, it is also worth noting that, assuming an adequate developmental scale exists, the Rasch analysis provides information about model misfit for persons as well as for items. Such person infit statistics can be highly diagnostic in subsequent uses of the STEP assessment system. This creates, for example, the capacity to identify distinct sub-set of children whose reading skill acquisition does not appear to follow the general pattern found in the overall scale. More detailed diagnostic workup and careful documentation of the progress (or failure to progress) of these children could help inform ongoing efforts to improve reading instruction.

Key to accomplishing all of this is the equating study design described above where items from different forms (yellow and purple) and step levels of the test are administered to different sub-samples of students. By organizing this common person equating across all steps and forms, the design allows estimation of the difficulty of each item in each step, the location of all item difficulties on one common scale, and the infit statistics necessary for examining scale adequacy.

### **An Empirical Test of the Developmental Scale**

Our first analysis of the construct validity of STEP examined whether the STEP items actually function as a single overall scale. Table 6 provides a summary of results on the infit statistics from the Rasch analysis of the STEP scale. As noted above, The Rasch model postulates a single underlying metric. Even with a perfect developmental scale, we expect under Rasch theory that 5 percent of the items will misfit purely by statistical chance. If the scale is working properly, the actual observed number of misfitting items should not grossly exceed this expectation. In fact, that is exactly what happened here. Overall, we found that out of the 591 items that comprise the total scale, only 5.1 percent had higher than expected mean squared errors.

**Table 6. Percent of Items Exceeding Expected Fit Values**

	Number of Items	% Above Expected
Overall Scale	591	5.1%
Sub-scales:		
LID (name and sounds)	80	7.5%
Phonological awareness	28	14.3%
Concepts about print	56	3.4%
Developmental Spelling	230	4.2%
Reading Accuracy	24	0.0%
Reading Rate	18	0.0%
Comprehension Questions	154	6.1%

In addition, we found no significant evidence of clustering among the “misfitting items” in any of the assessment sub-components. For example, 3.4 percent of the concepts about print items misfit and 4.2 percent of the developmental spelling items misfit.<sup>3</sup> The phonological awareness items, which include the subtasks of rhyming, identifying first sounds, and segmenting words into phonemes, are the only potential point of concern. Given, however that there are only 28 items in this sub-scale, the presence of four misfitting items is not statistically significant. Moreover, the fact that these misfitting items are spread across the three different assessment sub-tasks that comprise this sub-scale (one misfitting item in the rhyming sub-task; one in identifying first sounds, and two in segmenting words) suggests that the misfit may be related to a characteristic of the particular items rather than an indication that Phonological Awareness represents a separate underlying dimension.<sup>4</sup> In sum, the infit statistic results support a claim that STEP measures a single underlying developmental phenomenon.

### **A Theoretical Test of the Developmental Literacy Scale**

Next we considered whether the hierarchical ordering of the items within the STEP scale appear theoretical sensible given extant reading development theory. Under the Rasch

---

<sup>3</sup> We note that if the misfitting items clustered by sub-component, this would constitute evidence that this sub-component did not function consistent with the overall scale.

<sup>4</sup> For example, in the segmentation sub-task, students are asked to produces the phonemes that comprise the word, “grew.” This is the only word in the list with a consonant blend which may account for the misfit.

analysis, each item has an estimated difficulty which places it in a unique position on the overall scale. If STEP functions as a developmental scale, items should cluster within the scale by sub-component. That is, the Concepts about Print items, for example, should tend to be easier than the Phonological Awareness items which in turn should be easier than most of the reading accuracy ratings. Similarly, within each sub-component, the items difficulties should also order in theoretically predictable ways, e.g. the rhyming items within Phonological Awareness should generally be easier than items that examine a child's skill at segmentation.

**Ordering of the Assessment Sub-Components.** Figure 3 displays a box plot of the item difficulties for each assessment sub-component included in STEP. These boxplots highlight the range of item difficulties by sub-component and the relative ordering of the sub-component item clusters within the overall scale. In order to provide some concreteness in the discussion below, we have added to the display information about selected individual item difficulties and/or mean item difficulties for the sub-tasks within various the assessment components.<sup>5</sup>

Within the Rasch metric, both item difficulties and person measures are expressed in logits and have a direct relationship to one another. Any given person measure can be interpreted as a probabilistic statement about the likelihood that a child will answer correctly each item in the overall scale. For example, a student with a score of -4.00 on the Developmental Literacy Scale would have a .90 probability of answering correctly the average item within the 1-to-1 matching sub task in Concepts about Print (item difficulty/mean difficulty for 3 items -3.7). That same child would have over a 90 percent chance of knowing most letter names, and would very likely be also matching words with the same first sounds in Phonological Awareness. In contrast, we would not expect such a child to read with accuracy a Step 4 text (-1.2 item difficulty).

---

<sup>5</sup> Some of the item labels in Figure 3 represent the mean item difficulty for several items, for example upper-case letter names, rhyming, etc. For Comprehension, the labels for each step is the mean item difficulty for the comprehension questions at those steps.

As expected the easiest items within the STEP developmental literacy scale focus on concepts about print and letter names and sounds. Within Concepts about Print, the easiest item is “knowing that you read the text” (and not the pictures). Next students learn about directionality – reading left to right. The most difficult concept about print is understanding 1-to-1 matching of voice to words on the page.

Looking at the next boxplot, representing letter names and sounds, we see that as students are gain more understanding of Concept about Print, they are also learning to identify letters. Children tend to learn upper case letters first (-6.1 mean difficulty for 26 items for upper case; -5.2 mean difficulty for 28 lower case including print and type face fonts for ‘a’ and ‘g’) and to eventually associate letters with the sounds that they make (-3.0 mean difficulty for 26 items.)

The item difficulties for the Phonological Awareness sub-component (the third boxplot in Figure 3) indicate that these skills are developing in tandem with Concepts about Print and Letter Names and Sounds. Rhyming words comes first (-4.3 mean item difficulty) , following by matching words that begin with the same first sound, (-3.4 mean item difficulty) and finally learning to segment words into individual phonemes(-1.6 mean item difficulty). Note that being able to consistently segment words into phonemes does not occur until students are accurately reading text between Step 3 and Step 4.

Next comes the results for Text Reading Accuracy which follows a monotonic progression. The jumps in item difficulty across Steps 2 through 5 are relatively large. Students at this stage in learning to read are acquiring many new strategies and skills, and correspondingly, the ability to read accurately the next level text represents a significant developmental gain. A child with a scale score of -4.1 has a probability of 0.90 for reading a Step 2 text with mastery. To achieve a similar likelihood of mastery at Step 3 implies a scale score of -3.2; the corresponding jumps to reading accurately at Steps 4 and 5 translate into scale scores of -1.6 and -0.1 respectively. In contrast, for Step levels 5 and beyond, the distance between item difficulties become significantly smaller. Although students are continuing to learn new strategies and skills to decode more

complex texts, the actual amount of development occurring here appears less than at lower step levels (Carver 1995; Adams 1990).

Reading Rate item difficulties and Text Reading Accuracy levels display an interesting relationship with one another. As expected, the ability to read a text accurately precedes being able to read the same text with appropriate speed or rate. That is, the item difficulty associated with the reading rate at each step is significantly higher than the corresponding reading accuracy for that step. Of significance, these gaps become larger as students progress into reading the more complex texts involved in the upper levels of STEP. For example, the gap between accuracy and rate at Step 5 is approximately 0.5 logits on the scale. By Step 10, the gap has increased to 2.0 logits. These results are consistent with published research (Carver 1990; Carver & Leibert, 1995) that the ability to decode accurately more complex texts represents modest developmental gains, whereas to read such texts with fluency represents substantially greater improvement. This is highly significant because fluency in turn links most directly with comprehension.

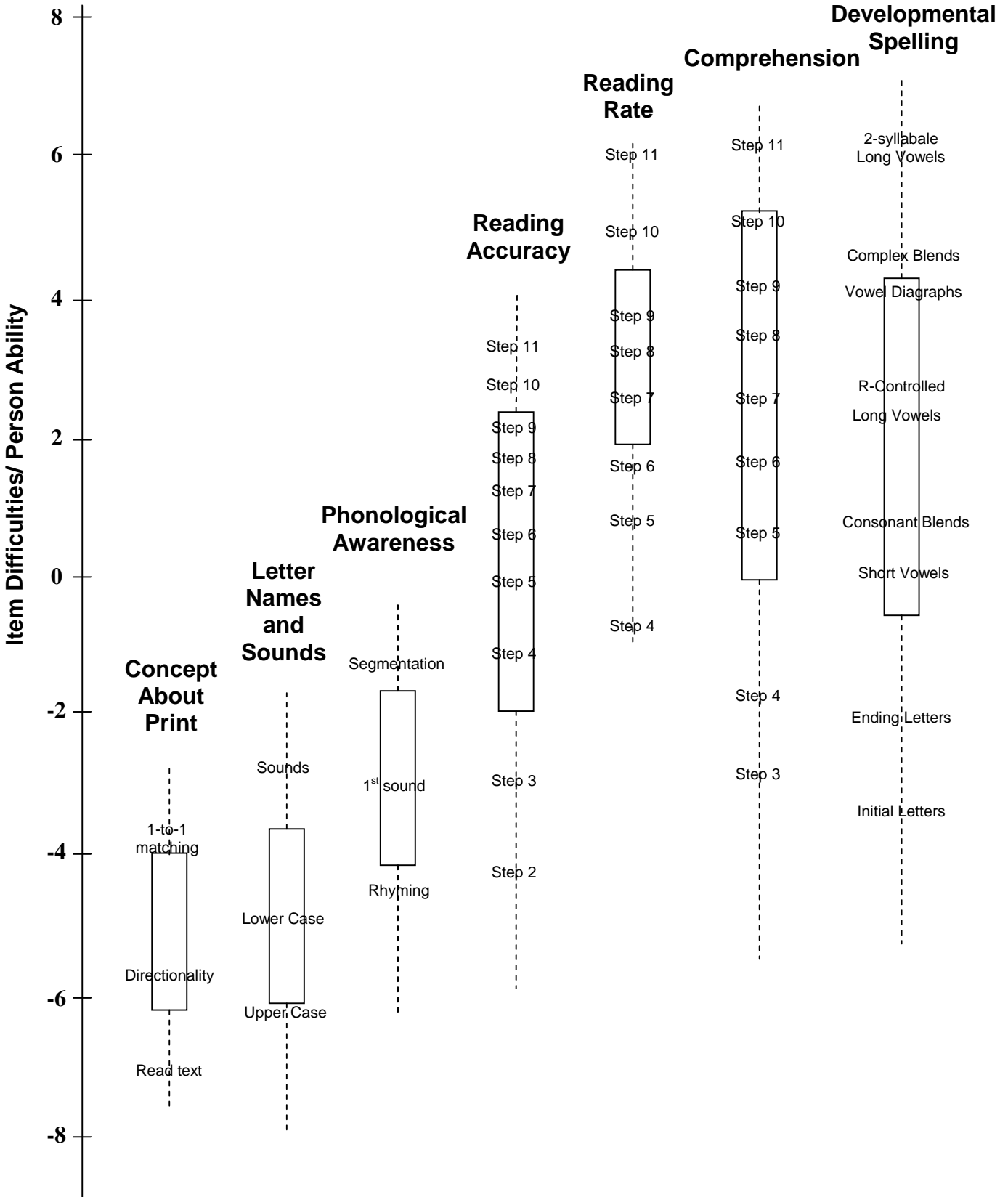
The boxplot for Comprehension displays the median item difficulty for the 5 to 8 comprehension questions associated with each text at each step level. Through Step 5, the difficulties for accuracy, rate and comprehension are fairly similar at each respective step. This implies that students at these levels are learning in tandem to read accurately, with a reasonable rate, and to comprehend these simple texts. As the text become more complex, however clear skill differentiation begins to emerge. As we saw with reading rate, a substantial gap appears at Step 6 and beyond that separates the comprehension item difficulties from a simple accurate decoding of text. To decode accurately these more complex texts represents an improvement, but the big developmental gains are associated with doing this with fluency and comprehension.

Finally, we see that items difficulties for Developmental Spelling also follow the expected pattern moving from writing initial and ending letters to short vowels and blends to long vowels and so on. (Henderson 1990; Bear, et. al. 2000; Ganske 2000). Moreover, these item results align in a reasonable fashion with the other sub-components

in STEP. For example, we see that students who are accurately reading text at Step 5 are also likely to be writing short vowels in developmental spelling (i.e., Step 5 accuracy and writing short vowels have approximately the same item difficulty.) Similarly, students who are reading a step 9 text accurately are spelling long vowel and r-controlled vowel patterns as well.

Taken overall, the item difficulty alignment of the sub-components within STEP (as well as the sub-tasks within each sub-component) is fully consistent with established scientific research on reading skill acquisition. These empirical results add credence to the construct validity of STEP as an integrated development scale.

Figure 3. Developmental Literacy Scale



**Ordering of the Items within the Assessment Sub-Components.** We now proceed to take a closer look at each of the sub-components that comprise the integrated STEP assessment system. This analysis provides further confirmation of the construct validity of STEP at a more micro level. Equally important, it elucidates how data on these items can help inform teachers about patterns in students' reading development in making sense of letters, words, and concepts.

*Concepts about Print.* Figure 4 displays the ordering of items within the Concepts about Print sub-scale. The data in the figure represents the mean difficulty of sub-sets of questions that tap the same concept as it appears in different texts within the STEP assessment system.<sup>6</sup> For example, students are asked to count the number of words on two different places in the nursery rhyme that the teacher is reading to the student. These items are indication of the students begin understand of what represents a word. Concepts such as reading the words on the page (versus the pictures), left-to-right directionality, and return sweep tend to be easiest and learned first. Items tapping these skills have scaled difficulties in the -6 to -7 logits range. This is consistent with published research that children typically learn the basics about how books work before focusing on words and letters (Morris 1993). In some respect, this may be described as “necessary” in order for students to understand the more advanced concepts about how letters and words function within the context of books.

Following this in order of increasing difficulty are questions which inquire about which letter is first and which is last in a word and asking children to count the number words on a line of text. This demonstrates an emerging understanding of the concept of a word (even before being able to read that word independently).

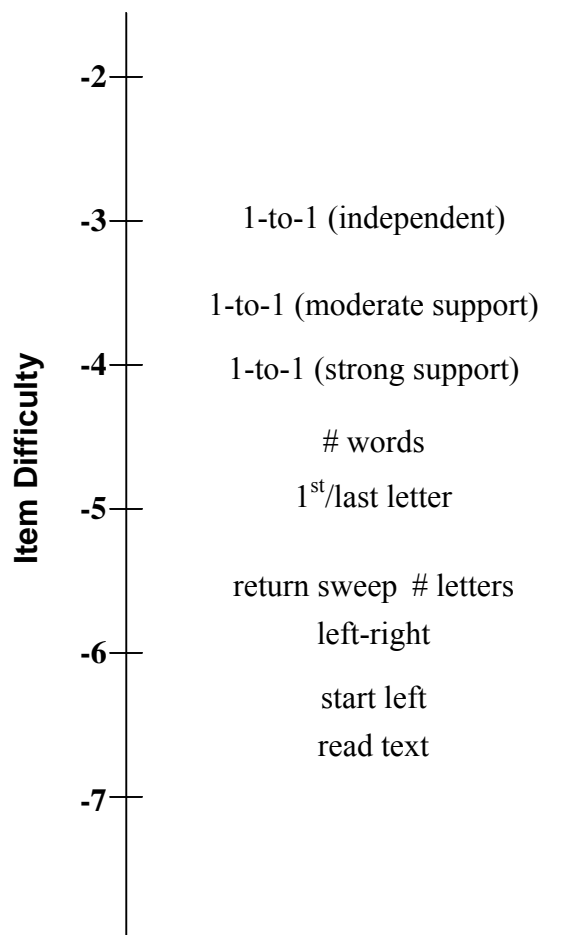
Finally, the most difficult items in this sub-scale tap into a developing sense of 1-to-1 matching. These items typically fall in the -3 to -4 logit difficult range. Three types of

---

<sup>6</sup> We note that some variance in item difficulty across books is observed though the overall pattern represented in the figure is consistent. There is simply more inherent measurement error at the lower end of the STEP scale because we are working with younger children (late pre-K and early kindergarten).

items assess a student’s growing understanding in this area. First, are items where the student is given “strong support” in the 1-to-1 matching task. The teacher reads a 3 to 5 word line from a nursery rhyme and models pointing at each word. Then, the teacher asks the student to point at the words while she reads it again. During 1-to-1 matching with “moderate support”, the teacher asks the child to point to the words on the first reading. One-to-one matching with “independence” is assessed using a pattern book. The teacher reads the repeating pattern in the book two times while modeling pointing to each word. Then, the student is asked to read the remainder of the book which contains this pattern and point at the words.

**Figure 4. Concepts about Print Item Map**



*Letter Identification and Letter Sounds.* The order in which children learn the names of letters depends on multiple factors such as letter shape, the frequency of exposure to particular letters in common environmental print, the frequency of letters as they appearance in commonly printed words and other factors as well. Nonetheless a reasonable overall pattern emerged in the ordering of the item difficulties for this sub-scale as well (See Figure 5).

Frequently encountered upper case letters (A, B) or letters with simple shapes (O, o, X, x) appear easiest for children to identify. Other letters that appear frequently in environmental print such as “M” (for MacDonalD’s) also anchor the bottom of this sub-scale in the -7 to -8 logit range.

The relationship between learning uppercase and lowercase letters appears to have two salient aspects. Lower case letters with similar shapes as the upper case letters are learned at about the same time, that is, they have relative similar item difficulties, for example, the pairs X with x, S with s, C with c, and K and k all have item difficulties with 0.1 logits respectively of each other. Lower case letters that have different shapes from their upper case equivalent tend to be learned after their corresponding upper case form. The item difficulties here differ by .5 to 1.0 logits; compare in Figure 5, for example, the item pairs, H with h, N with n, T with t, F with f, and L with l.

The letter name assessment also includes the two forms of the lowercase letter “a” and “g.” The “hand printed” versions (a, g) are recognized earlier than their book print form equivalents (a, g). This makes sense as most children are first introduced to the “hand printed” version and later encounter it only in books.

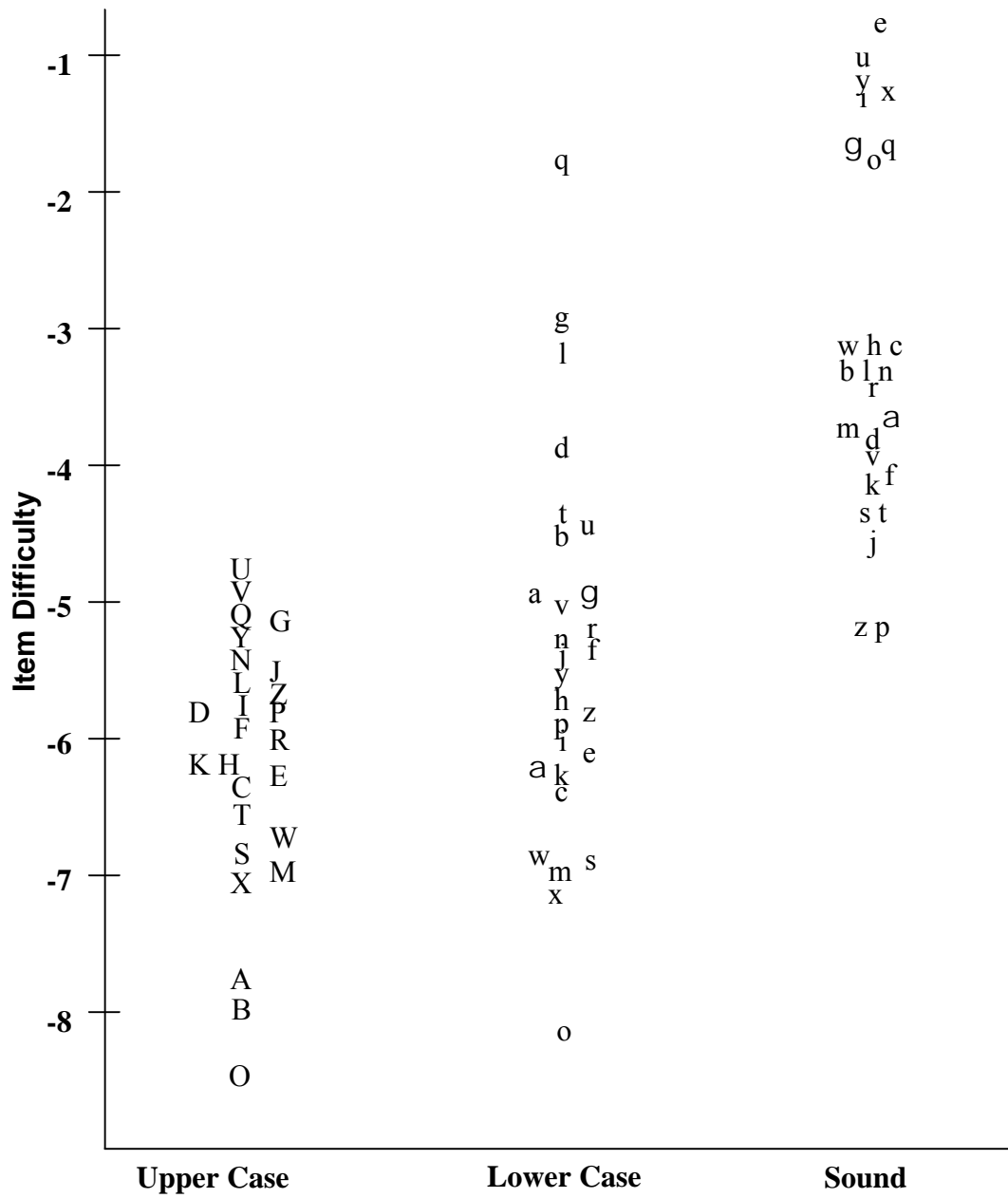
The letters b, d, and q are among the most difficult letters for a child to name, likely because of their similar shape – b is a reversed d and q a reversed p. In addition, the letter q is not frequently encountered in early level books, adding further to its placement near the top end of this sub-scale with estimated item difficulties in the range of -4 to -2 logits.

In terms of the Letter Sounds task within this sub-scale, these skills are typically acquired after students have learned the particular letter name. That is, a student has to consistently recognize the letter symbol before he or she reliably associates a sound with that letter. Knowing the name of the letter, in essence, creates an anchor for further understanding how the letter functions in other contexts (Treiman et al. 1998; McBride-Chang 1999).

Letters whose sounds are similar to their names tend to be learned first, notice for example, that z, p, t, and k are at the low end of the difficulty scale for the items within this sub-task. This may in part be a function of the nature of the STEP assessment task itself as the letter names and letter sounds are being asked for in isolation rather than in the context of words. However, it is also likely related to a conceptual shift in students' understanding that separates the letter name from how it functions in words (Treiman et al. 1998).

Finally, the production of short vowel sounds is secured last. The items anchor the top of this sub-scale with difficulties in the range of -2 to -1 logits. Within this set of short vowels, we note that a common pattern occurs both here as well as in the developmental spelling sub-scale (see details) below. In both cases, the short /a/ is learned first and /e/ is acquired last?

**Figure 5. Letter Identification and Letter Sound Item Map**



*Phonemic Awareness.* Figure 6 displays the relationship among the items within the three phonological awareness tasks that comprise this sub-scale. Each item is labeled with a number to identify its order within the assessment and with the specific the word used as the assessment prompt. For example, under rhyming, “7-clock” was the seventh item in a task that asks students to match the word “clock” (which is represented by a picture in the assessment) with a corresponding word that rhymes with it. This had an item difficulty of -4.6. Similarly, with Matching First Sound, “8-duck” was the eighth item and asked the student to point to the picture that begins with the same sound as “duck.” The item difficulty for this task was -3.8.

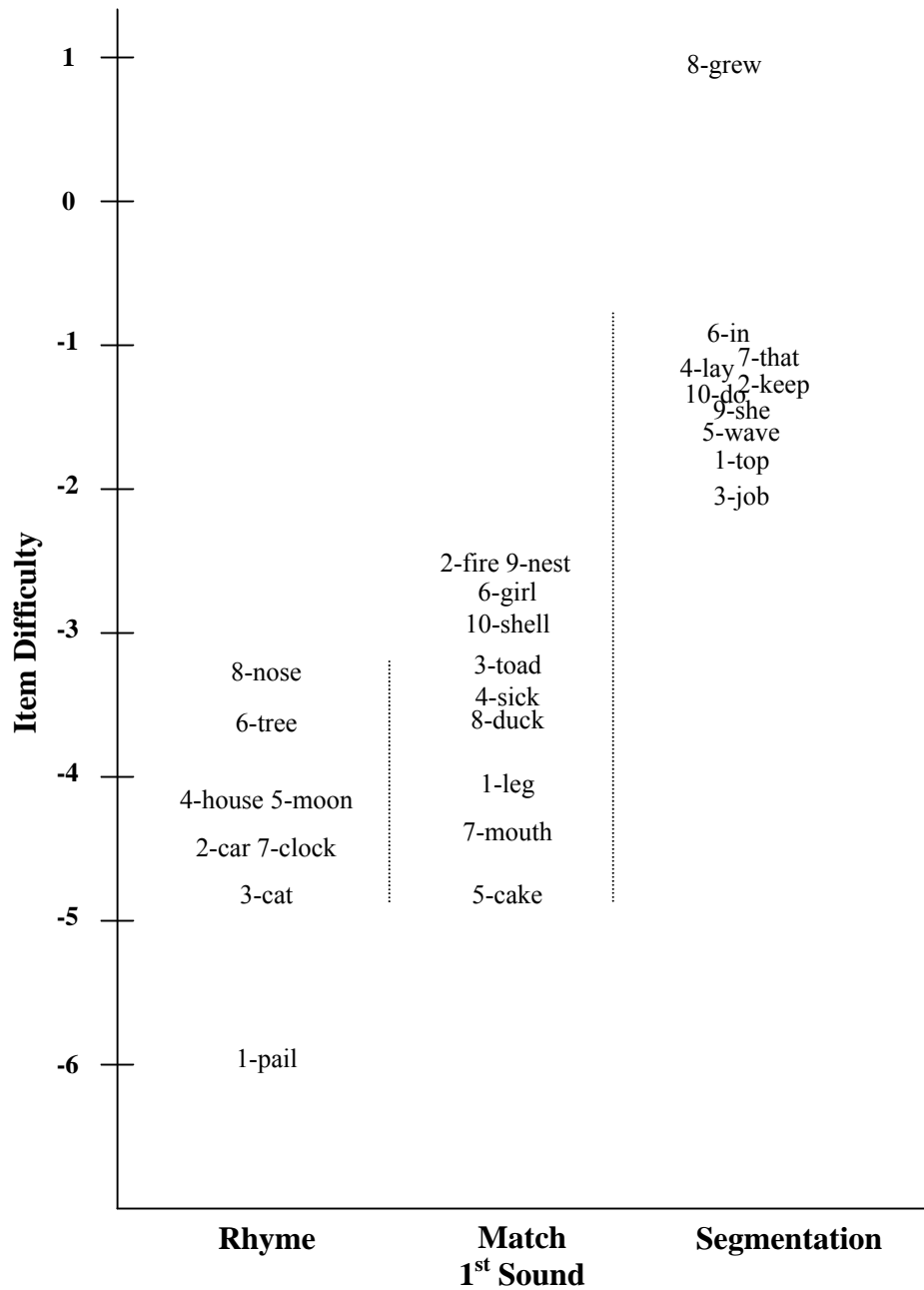
The ordering of the item difficulties for this sub-scale displays strong coherence as an overall set. These results are consistent with the idea of a single underlying ability of increasing student sensitivity to linguistic units – from words that have the same ending sound unit in rhyming, to matching initial sounds which requires separating first sounds in words, to segmenting all the phonemes of words. (See Anthony and Lonigan 2004 for further details on these relationships.)

In general, rhyming is the easiest task. The mean difficulty for this item subset is -4.3 logits. This is followed by matching first sounds with a mean item difficulty of -3.4 logits. Notice, however, that there is considerable overlap in the item difficulty of rhyming and matching first sounds. Segmenting the sounds in two- to three-phoneme words represents the most difficult task set (a mean item difficulty of -1.6). Words with a C-V-C pattern (job, top) are the easiest in this task. These have item difficulties around -2 logits. The difficulties for the other two- and three-phoneme words in this assessment task cluster closely together in the -1 to -2 logit range. “Grew” is the only word with a consonant blend and is significantly more difficult to segment than the other words on the list.<sup>7</sup>

---

<sup>7</sup> The segmentation task is scored on segmenting the entire word correctly. An alternative would be to score the segmenting of the individual phonemes so that words could receive partial credit. For example, a student could respond to “wave” by segmenting onset and rime (/w/ /av/). Partial credit score may provide further diagnostic depth. This alternative is currently being explored.

**Figure 6. Phonological Awareness Item Map**



*Accuracy, Reading Rate, and Comprehension.* Reading words in connect text, reading with sufficient speed, and drawing meaning from the text are intrinsically linked. However, it is important to explore how these relationships may shift as students encounter more complex books. In general, these three aspects of reading tend to develop together -- more accurate readers tend to read quicker and with more comprehension (Pinnell, et al., 1995). Similarly, if reading accuracy declines as when students encounter higher level books, their reading rates as well as comprehension also declines (Carver 1990). Nevertheless, many children are able to accurately decode words in higher level texts while not understanding the concepts or plots they are reading. That is, accuracy appears to be a necessary though not sufficient condition for comprehension.

In addition, reading rate (as well as fluency) plays an important mediating role between accuracy and understanding. Readers have a limited amount of attention for any given cognitive task (LaBerge & Samuels 1974; Perfetti 1985; Stanovich 1980). Students who expend more time in problem-solving words have less cognitive space to make important connections in the text to support their understanding. Students' reading rates are indicative of this "trade-off" between decoding focus and comprehension (Adams 1990).

Thus, looking at their interrelationship of accuracy, reading rate, and comprehension across the text levels will serve to further explicate the developmental reading process. The developmental literacy scale provides empirical evidence of these interrelationships as well as direct information about the construct validity of the assessment.

*Development of Leveled Texts.* A set of leveled text is at the core of this part of the assessment. Their careful development and evaluation is therefore crucial. Fountas and Pinnell’s descriptive text leveling system (described in section above) formed the basis for writing the texts. Authors were provided with descriptions of each level as well as several benchmark examples of published books. This formed the basis of first drafts of each book.

The newly written books were also compared with the Reading Recovery texts used in the early pilot to establish comparability. Because the Reading Recovery books were limited in their topics and range, adding new books with more variety was essential in extending the assessment.

The next phase of development was to perform Lexile analysis of the books in each series. The method takes into account word frequency and sentence length as the key indicators to measure readability (Stenner 1996). This served as a baseline reference for initial revision. Earlier versions of the books were edited in terms of word selection and sentence length based on the analysis. Table 6 represents the final Lexile levels after this revision.<sup>8</sup>

**Table 6. Lexile Analysis of Leveled Texts**

	Yellow Series	Purple Series
Step 5	279	276
Step 6	311	321
Step 7	349	352
Step 8	393	408
Step 9	437	442
Step 10	498	482
Step 11	542	557
Step 12	598	589

The results show a relatively smooth progress in estimated text difficulty. The ultimate question about the validity of the text levels, however, has to be evaluated in the context

<sup>8</sup> Lower text levels are much more strongly related to picture support. Lexile rating does not take this into account and is, thus, not appropriate for analysis.

of how children actually respond to the books in terms of decoding, reading rate, and comprehension. This direct analysis is detailed below.

*Reading Accuracy.* Focusing first on reading accuracy, Figure 7 shows the relative difficulty of reading the text at each Step with 90 percent accuracy. In parallel with the Lexile analysis, the text levels monotonically increase in difficulty. However, the difference between texts is not evenly distributed. In general, a larger distance in difficulty is seen between texts at the lower levels than between those at higher levels.

The largest “distance” is from Steps 3 to 4 (over 1.75 logits). The key factor determining movement at this level is problem solving the words of the text. It is at this point in early reading that students are learning enormous amounts about how letter patterns function and how to use this information to solve words. Thus, as anticipated, the scale shows students who are able to move from reading a Step 3 book to a Step 4 book accurately are making large gains.<sup>9</sup>

As we progress to higher text levels, the additional demand to reach decoding accuracy begins to decrease. The difference in item difficulty between Steps 4-7 is on average .75 logits; between Steps 7-11, the average difference between Steps reduces to .52 logits. This suggests that the pivotal skill of problem-solving words and reading accurately becomes less of a hurdle as text levels increase.

*Reading Rate.* The center column of Figure 7 displays the relative difficulty for the reading rate targets at each Step. The reading rate target increases across Steps. (The target in words per minute (w/m) is shown in parentheses next to each Step.) The targets were set based on a review of the literature (Carver 1990, Pinnell, et al., 1995) as well as theoretical considerations. As students begin to read stories with more complex ideas and

---

<sup>9</sup> It should be noted that it may be possible to write additional text that fall between Step 3 and Step 4 in difficulty. However, such fine-grained, formal assessment of text was not chosen because the information acquired is intended for classroom teachers. These smaller distinctions may prove very useful for one-on-one tutoring (such as Reading Recovery) but for thinking about instruction for small groups or whole classrooms such detail may be overwhelming.

concepts, they will need to give proportionally more cognitive attention to making meaning while they read (rather than to problem-solving words).

Consequently, the item difficulties for reading rates are influenced by two factors. First, students are reading text with more new vocabulary as well as longer sentences. Therefore, for example, we would expect that reading a Step 5 text at 40 words per minute will be easier than reading a Step 6 text at the same rate. The item difficulties on the developmental scale confirm this. For Step 5, the difficulty is .85 logits; for Step 6, 1.62 logits.

Second, because the reading rate targets also increase across the Steps, the rising target will add a further increase in difficulty. This pattern is also evident on the scale. Item difficulties increase the most where the target reading rates were increased at for example Steps 4 to 5 and Steps 9 to 10.

*Comprehension.* Figure 7 also displays the average item difficulty for comprehension questions at each Step. The number of questions for each Step ranges from five to eight.<sup>10</sup> As with accuracy and reading rate, the item difficulties increase monotonically. Again, the results provide more evidence for the appropriate leveling of the text.

Several interesting patterns emerge across the Steps in relationship to comprehension. The difference in average item difficulty between Step 3 and Step 4 (.49 logits) is relatively small in comparison to the difference in reading accuracy difficulty (1.75 logits). This suggests that moving from Step 3 to Step 4 is more dependent on being able to problem-solve the words of the text than comprehending the book. It is important to note that illustrations are a strong support to comprehension at these levels. In contrast, a significantly larger difference in comprehension between Step 4 and Step 5 is evident for comprehension (2.21 logits). It is at this level that picture support is not sufficient to

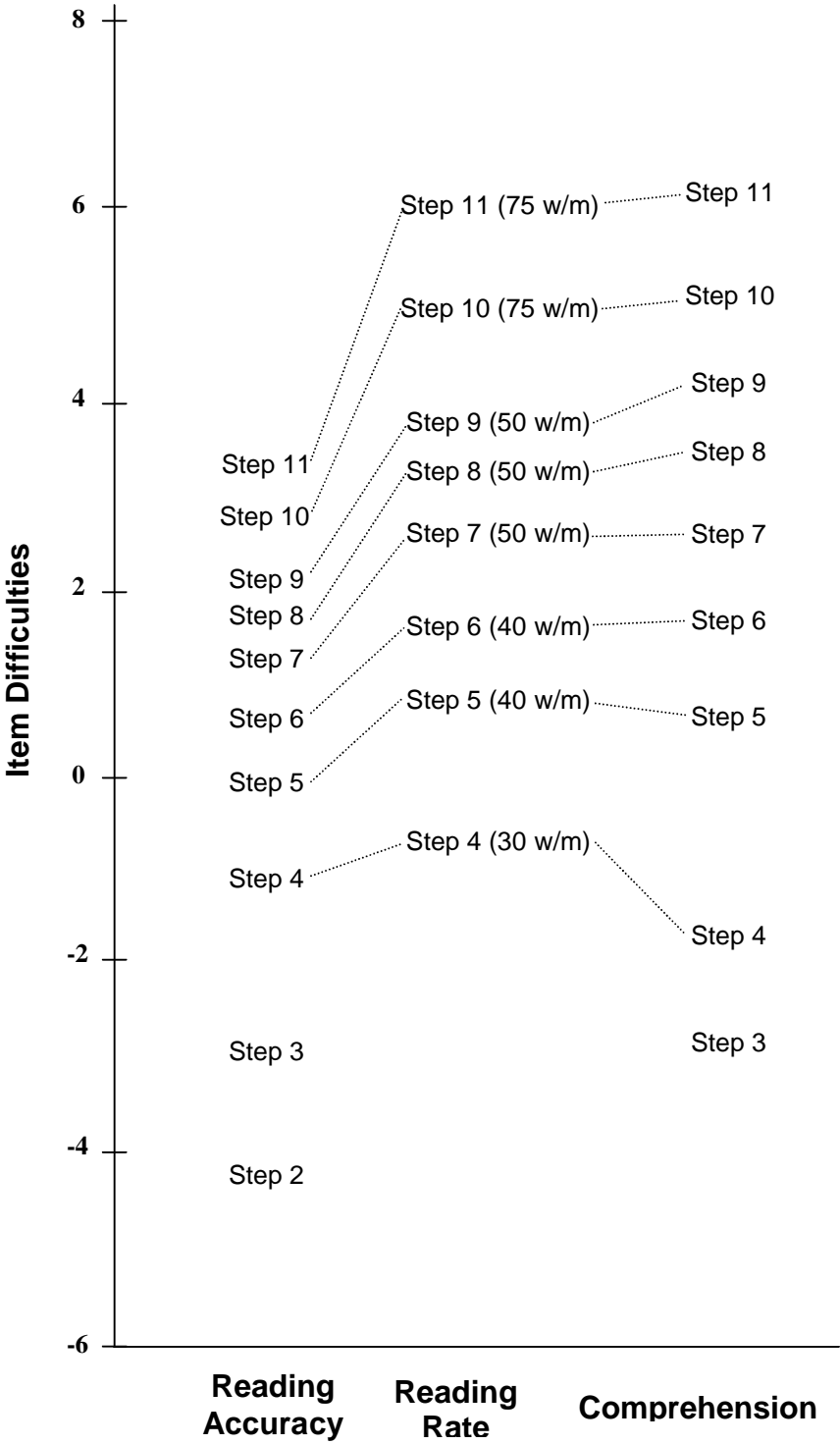
---

<sup>10</sup> The item difficulties for a limited number of questions (11 of 134) were more than 1.5 logits from the next closest items on that Step. These items were deleted for the calculation of the Step means. We have begun work to rewrite these questions or to substitute new questions.

promote detailed understanding of the story. Inferential thinking at this level depends on drawing meaning from the text (and pictures become secondary). This is a significantly more challenging task. The comparative difference in item difficulties between Step 4 and Step 5 in reading accuracy is .86 logits. Thus, moving from Step 4 to Step 5 requires a comparatively large increase in comprehension ability.

Finally, from Step 5 to Step 11, increases in average comprehension difficulties are relatively constant – approximately 1 logit for each Step. Building comprehension strategies as text become more complex plays a crucial role across the entire range of book levels. The difficulty of reaching the target for reading accuracy for the same Step lags further and further behind. At Step 5, the difference is .52 logits. This shifts to 1.15 logits by Step 7 and to 2.39 logits by Step 11. This trend confirms the increasingly prominent role of comprehension as students move into higher level books.

**Figure 7. Accuracy, Rate, and Comprehension Item Difficulties**



*Developmental Spelling.* Developmental spelling theory suggests that invented spelling is a window into a student’s knowledge of how written words work and can be used to guide instruction (Invernizzi, Abouzeid, & Gill, 1994). Correct spelling of certain word features and specific kinds of spelling errors reflect a progressive differentiation of word elements by students which determine how words are *read* as well as written (Bear, 1991; Bear & Templeton 1998).

Assessment from this perspective focuses on spelling features of words instead of considering only the correct spelling of whole words. It is students’ knowledge of spelling features that seems to be more related to reading ability (see Ehri 1997). Spelling as well as reading advance in stage-like progression, sharing important conceptual dimensions.

As described in the previous sections (see Table2), this progression is characterized by key representations of letter patterns: starting with beginning and ending consonants, moving to short vowels and consonant blends, and then to long vowel patterns, r-controlled vowels, and diphthongs. As students acquire correct use of the more simple patterns, they begin to “use but confuse” those that come immediately after while the more advanced feature may be absent altogether. For example, a student may use “fep” for “flip.” In this case, she has beginning and ending consonants under control and is representing short vowels although not yet distinguishing them.

Figure 8 is organized around these letter patterns and ordered by the Step in which they are featured. Each column displays the range of words highlighting a particular letter pattern. The particular letter pattern is demarcated in the word with capital letters (e.g., the consonant blend “pl” in “PLum”).

The results confirm the theoretical development pattern described in the literature. First consonants are clearly learned first with most items clustering around -3.75 logits. Some ending consonant are also being written correctly at this level (such at “poT”) but most

are not independently written correctly until later. The median item difficulty centers around -2.75 between “jeT” and “siP.”

Short vowels are assessed both in C-V-C words as well as words with blends or diagraphs. Correctly representing short vowels appear much later than beginning and ending sounds with an item difficulty centering around -.5 logits. Short /a/ tends to be the easiest. It is interesting to note (as shown in Figure 5) that short /a/ is also the first vowel sound that students are able to produce in isolation. Other short vowel sounds are produced in isolation with item difficulties around -1.10 logits. Thus, knowing and producing the sound in isolation precede their correct use in writing words.

Consonant blends (e.g., “pl,” “sm”) and diagraphs (e.g., “sh,” “th”) are at approximately the same level as short vowels indicating that they tend to be developing at essentially the same stage of development. However, ending blends such as “luMP” and “reNT” tend to be more difficult. The nasal quality of /m/ and /n/ causes these sounds to be overshadowed by the consonant that follows, making it more likely for them to be overlooked (Ganske 2000).

The next set of word features assessed at Steps 6-7 and Steps 8-10 are less ordered in their appearance. Long vowel patterns (e.g., V-C-e, “ai”, “ee”) are interspersed with r-controlled (e.g., dARk, shirt) in terms of their item difficulties. This suggests that these aspects tend to be learned in tandem. One does not necessarily precede the other. However, both of these word features are much more difficult than the more simple consonant blends and diagraphs.

Vowel diphthongs (stOOd, pOInt) as well as complex blends (paTCH, STReam) are somewhat more difficult with most appearing between 3.00 and 4.00 logits on the scale. Despite the pattern complexities in these words, as students read and examine words, they gradually sort out the correct use of patterns in single-syllable words.

The final letter patterns involve two-syllable words and extend previous learned patterns to these longer words. Students are required to think about double consonants at syllable junctures both when adding endings (shaKING, baGGED) and in the middle of words (teNNis, baTTer). In addition, long vowels and r-controlled vowels are now embedded in syllable rather than entire words (retAIn, wARning). This adds another layer of complexity and, thus, these patterns tend to be more difficult than their single syllable counterparts.

In sum, the analysis of item difficulties in developmental spelling strongly confirms the validity of the assessment and provides an empirical base of evidence for developmental spelling theory.

**Figure 8. Developmental Spelling Item Map**

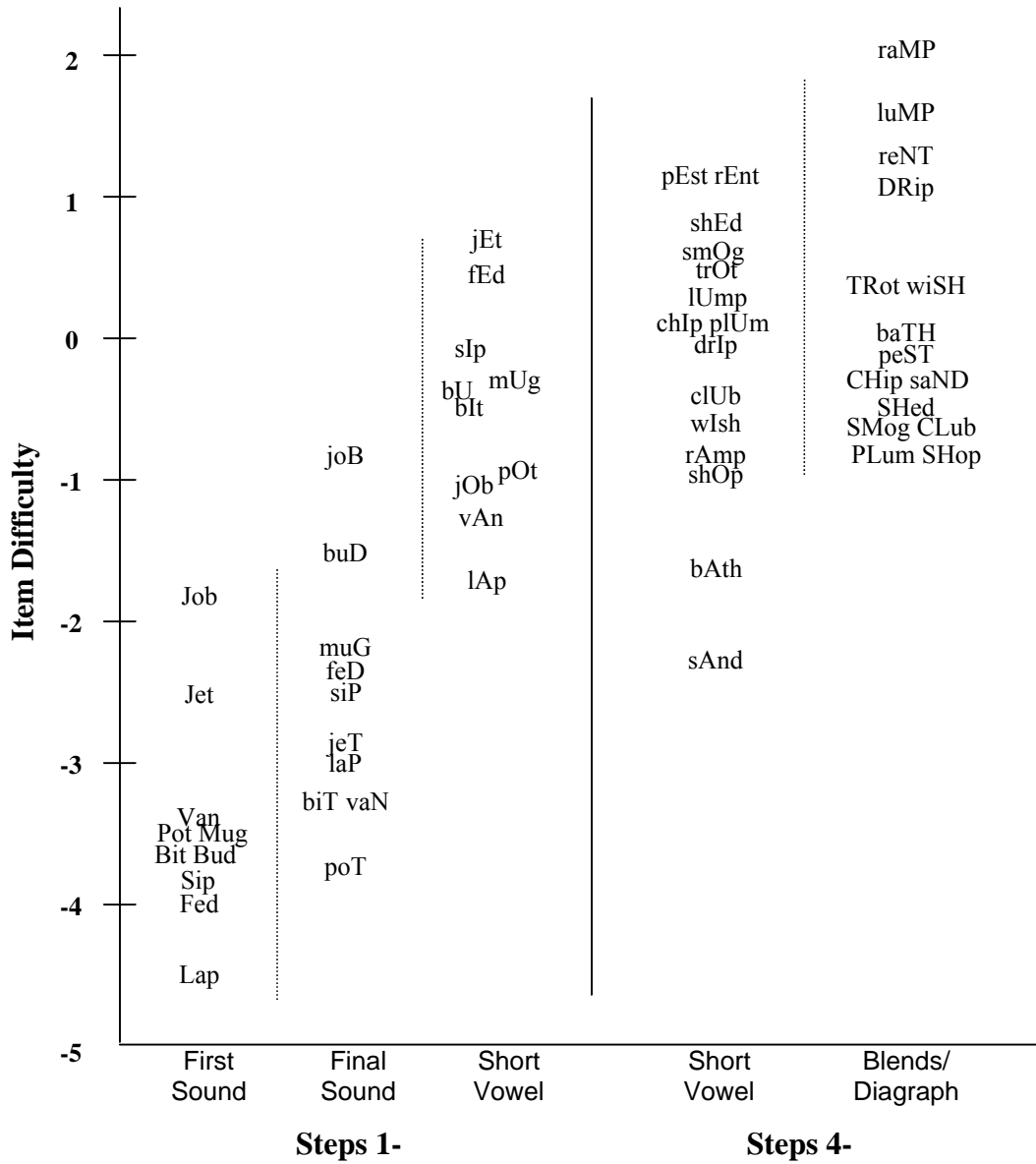
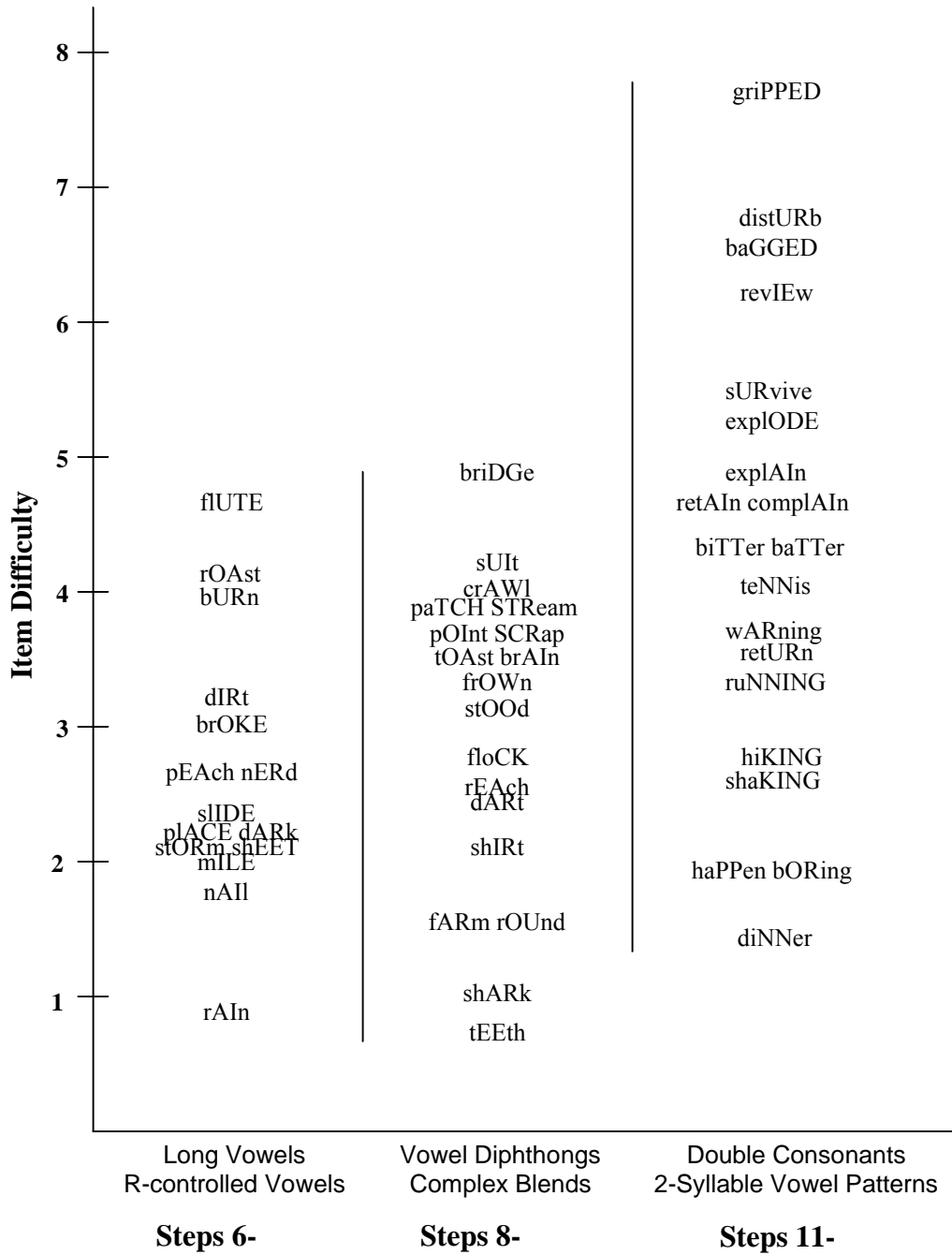


Figure 8 continued



### **Internal Reliability**

STEP is designed to be used as a classroom-based assessment. It is intended for use by teachers and other school staff in course of their daily practice. Therefore, we sought to establish reliability-in-use by collecting assessments administered by regular teachers in ordinary classrooms.

Typical practice in schools consists of administering STEP assessments to students at three to four time points during the academic year. At each time point, students may be assessed at two to three step levels depending upon their progress from the previous time point. Teachers are able to compare a full range of data from one time point to the next, including results from earlier assessments. This expands the information available to teachers for purposes of making reliable judgments about each individual child's development as a reader.

The reliability study results reported below can best be described as an analysis of "reliability in the context of regular school use" in contrast to reliability established under more scientifically-controlled research settings. As detailed earlier, (See section on "Description of Psychometric Studies, Construct Validity-Internal Reliability Study") all of the assessment data analyzed below were gathered by regular teachers and classroom assistants. While all had received the basic training in STEP administration, data collection occurred in regular classrooms within the context of normal day-to-day instruction.<sup>11</sup> No special "testing conditions" were imposed for purposes of this study. As a result, we view the results reported below as lower bound estimates given the highly variable conditions typically found in the disadvantaged urban school classrooms where these data were collected. If STEP were used in scientific studies under more standardized administration conditions (e.g. a quiet, secured space where test

---

<sup>11</sup> Training involved at least two four-hour sessions on administration and interpretation of results. Workshops included practice with video as well as teacher administration of the STEP to two to three students in their own classroom which were discussed in a follow-up session.

administration would be unlikely to be interrupted by other classroom events), we would expect even more reliable results.

### **Conceptualizing Reliability in the Context of School Use**

Reliability is a property of a measurement instrument or assessment procedure applied to some population of individuals. The “population” may take on different definitions depending upon the particular context of use for any instrument. For purposes of summarizing our results, we consider below three different contexts of use typical in school settings:

1. Making distinctions among children’s state of reading development from kindergarten through grade 3. This “*Overall Scale Reliability*” is informative when using STEP data for informing program accountability questions (e.g. are we seeing improvements over time in our K-3 reading program?) It also provides relevant information for instrumentation decisions in scientific studies of reading development over this same set of grades or age period.
2. Making distinctions among children who are at same grade level. This “*Grade Level Reliability*” is a key consideration for example when school staff seek to make decisions, based on STEP data, as to which first grade students are “at risk” and should be assigned to supplemental reading services.
3. Making distinctions among small sub-groups of children who are at approximately the same developmental level for purposes of guiding strategic teaching in reading groups and targeted mini-lessons. We refer to the latter as “*STEP-level Diagnostic Reliability.*” This refers to the capacity of the instrument to inform teachers’ more micro-level instructional decision making within their classrooms.

Clearly, as one moves down this cascade from overall scale use to within classroom diagnosis, we increase the information demands placed on a scale.

## **Overall Scale Reliability**

**Full scale.** Table 7 summarizes the results from our Rasch Analysis of the K-3 student STEP data. It provides information on both person and item-level reliability statistics for the entire scale. Of primary concern is the “Person Reliability based on Real Root Mean Square Errors.” This statistic within a Rasch analysis is equivalent to a Cronbach’s alpha reliability coefficient (insert cite, ask Stuart). The overall scale reliability coefficient is 0.98 which implies that STEP has a very high degree of precision in distinguishing among students K to 3 in their developmental reading state.

Table 1 also includes information of item-level reliability, i.e. how precisely we have established the difficulty estimates for each item. Since STEP is routinely used as a developmental assessment, with students at different developmental levels taking different sub-set of items, having precise item difficulty estimates is important in establishing the vertical equating necessary for reliable score comparisons across the full scale. These item reliabilities are also very high (0.98).

**Scale sub-components.** As noted earlier, the STEP assessment system is built around five sub-components: Letter Identification (items include letter names and sounds in Pre-Reading to Step 3), Concepts about Print (Pre-Reading to Step 1), Phonemic Awareness (Pre-Reading to Step 3), Developmental Spelling (Steps 1 to 12), and Text Reading, including accuracy, rate, and comprehension (Steps 2 to 12). Since for some analytic and research purposes we might wish to consider students’ reading development by component, we also examined the scale reliability separately for each major component.

For Letter Identification, there is high sub-scale reliability (0.96) as well as high reliability for each of the two tasks, letter naming (0.95) and letter sounds (0.85), which combine to comprise this sub-scale. The sub-scale reliability for Concepts about Print was 0.72. The Phonemic Awareness sub-scale, which consists of three distinct tasks: rhyming, matching, first sounds and segmentation, has a slightly higher reliability of

0.77. Of the separate component tasks that comprise this sub-scale, only segmentation with reliability of 0.60 based on 10 items evidenced adequate reliability at the specific task level. Neither the rhyming nor matching first sounds task displayed adequate reliability at this level in our data set.<sup>12</sup>

---

<sup>12</sup> These results merit a bit of further elaboration. All of the item difficulties associated with these tasks are well estimated and there was no evidence of significant item misfit which would have indicated possible conceptual problems in the overall scaling. Rather, the low task reliability here is associated with the fact that we did not have a full range of student abilities present, at the task-level, in our psychometric study. For example, on the Rhyming task 53% of the children included in the Construct Validity-Internal Reliability Study scored either a 7 or 8 items correctly out of maximum possible score of 8. For first sounds, 61% scored 9 or 10 items correct out of a maximum of 10. With greater population variability at this specific task level, it is quite likely that the task scores would have been reliable.

**Table 7. Rasch Analysis: Person and Item Reliability**

## Person Reliability Statistics

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	
MEAN	41.5	61.5	-0.53	0.48	0.97	-0.2	1.06	-0.1	
S.D.	29.5	28.8	4.14	0.21	0.23	1.0	1.21	0.9	
REAL RMSE	0.54	ADJ.SD	4.10	SEPARATION	7.53	PERSON RELIABILITY	0.98		
MODEL RMSE	0.53	ADJ.SD	4.11	SEPARATION	7.81	PERSON RELIABILITY	0.98		
S.E. OF PERSON MEAN	0.13								
WITH	71	EXTREME PERSONS	=	1135	PERSONS	MEAN	-0.64	S.D.	4.22
REAL RMSE	0.64	ADJ.SD	4.17	SEPARATION	6.49	PERSON RELIABILITY	0.98		
MODEL RMSE	0.63	ADJ.SD	4.17	SEPARATION	6.65	PERSON RELIABILITY	0.98		

## Item Reliability Statistics

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	
MEAN	80.3	119.0	0.00	0.45	0.99	-0.2	1.07	0.0	
S.D.	88.1	128.7	3.89	0.31	0.32	1.6	0.93	1.3	
REAL RMSE	0.58	ADJ.SD	3.85	SEPARATION	6.62	ITEM RELIABILITY	0.98		
MODEL RMSE	0.54	ADJ.SD	3.86	SEPARATION	7.10	ITEM RELIABILITY	0.98		
S.E. OF ITEM MEAN	0.17								
WITH	41	EXTREME ITEMS	=	591	ITEMS	MEAN	-0.06	S.D.	4.04
REAL RMSE	0.71	ADJ.SD	3.97	SEPARATION	5.61	ITEM RELIABILITY	0.97		
MODEL RMSE	0.68	ADJ.SD	3.98	SEPARATION	5.85	ITEM RELIABILITY	0.97		

The sub-scale reliability for Developmental Spelling is 0.97. Since a common word list is used across a cluster of adjacent steps, we also investigated the reliability at this micro-task level (i.e. the ability to spell each separate word list.) For the word list used in Steps 1-3, the reliability was 0.75. Correspondingly, the reliability was 0.86 for Steps 4-5, 0.81 for Steps 6-7, 0.84 for Steps 8-10, and 0.88 for the common word list used at Steps 11-12.

For text reading, the overall sub-scale reliability is 0.82. This is based on a composite score that combines information on accuracy, reading rate and students' responses to the short list of comprehension questions that accompanies each story read. The reliability for this sub-scale drops to 0.70 if only the accuracy and reading rate tasks are used.<sup>13</sup>

### **Grade-Level Reliability**

For purposes of estimating STEP reliability at each grade level (K through 3), we divided the data from the Construct Validity-Internal Reliability Study into separate subsets by grade level (See Table 8). We then considered student results from a single test administration in late spring of 2003. Following standard STEP guidelines, teachers had administered assessments at 2 to 3 step levels beginning at the step thought most appropriate for each student based on teachers' classroom observations and prior data if available. Depending upon student results on the initial step assessed, teachers would either move up to the next higher level or back to an easier level. Within each grade, a span of 6 to 7 step levels is represented. This simply reflects the relatively wide variability in student development that exist within any given primary grade.

---

<sup>13</sup> We note that the design of the overall STEP assessment system is not optimal if the primary goal is to extract a separate reading accuracy and rate score for each child. Under standard STEP test administration guidelines, even if students read a passage with speed and accuracy, they are not exposed to the next higher level reading task, if they failed to answer correctly the comprehension questions associated with the current task. If one's interest is precise information about reading accuracy and rate, then a modification in this test administration procedure would be warranted. Under such a modified test administration procedure, the reliability estimates would likely be higher than those reported above since students would be exposed to more texts than is the case in a standard STEP administration.

In order to estimate the person level reliability (Cronbach's alpha), we ran a separate Rasch analysis on each grade level data set. Table 2 presents these results. These grade level reliabilities are quite high exceeding .90 at every grade. This means that STEP provides highly reliable data for informing decisions about individual students within classrooms such as placement into reading groups and assignment to supplemental reading services.

**Table 8. Person Reliabilities for Each Grade Levels**

	Step span administered at each grade level	Person Reliability (Cronbach alpha)
Kindergarten	Pre-reading to step 5	.98
Grade 1	Steps 2 -7	.95
Grade 2	Steps 4-10	.96
Grade 3	Steps 6-12	.93

### **Step-level Diagnostic Reliability**

Finally, we also examined the reliability of the STEP assessment system at a more micro level consistent with how STEP information might typically be used by teachers within classrooms. Here we examine the ability of the instrument to discriminate possible task-level differences in performance for students who are thought to be at the same general developmental level. For purposes of this reliability study, we grouped individuals together based on the highest step level that they had achieved in the immediately prior test administration, approximately 10 weeks earlier, and then proceeded to administer to each group the next 2 step levels respectively. The reliability associated with each two-step combination appears in Table 9. Clearly, STEP displays excellent properties at this level as well with reliability estimates for most of the two-step combinations hovering around .90. These results indicates that STEP data are capable of informing relatively fine-grain teacher decision making based on observable differences in individual student performance at any given step level.

**Table 9. Person Reliabilities for Adjacent Steps**

Steps administered	reliability	# of Items
Pre-R –Step 1	.96	154
Step 1-2	.96	150
Step 2-3	.96	132
Step 3-4	.91	110
Step 4-5	.87	46
Step 5-6	.92	63
Step 6-7	.83	35
Step 7-8	.87	55
Step 8-9	.84	40
Step 9-10	.84	40
Step 10-11	.88	60

### Concurrent and Predictive Validity

The concurrent and predictive validity results reported on in this section are drawn from the Concurrent-Predictive Validity Study described earlier. For purposes of examining the validity of STEP classroom assessments, we compared students' end of year STEP scores at grades 2 and 3 to their results on two standardized tests routinely used in the Chicago Public Schools (CPS). The CPS administered annually the Iowa Tests of Basic Skills in reading at both of these grades (Hover, et al., 2001). In addition, at grade 3 the state of Illinois administers its own standardized reading assessment, the Illinois Standards Achievement Test, which has been designed to align with Illinois state reading standards (Illinois State Board of Education, 2004). We compared our classroom-based STEP results to both of these external assessments since Chicago schools are accountable for student performance on both of these indicators.

It is also important to remember that the STEP data were collected under regular day-to-day classroom conditions which are much less controlled than those mandated for standardized ITBS and ISAT administrations. Thus, we view the results presented below as lower bound estimates of validity coefficients. Stronger relations would likely have emerged had we chosen to administer STEP under the same controlled conditions as

routinely used with standardized accountability testing. Also, in the spirit of focusing on validity in the context of classroom use of STEP, all of the results below are based on the recorded step levels for children rather than the more fine-grained scale score. While the latter would be the focus in research applications (and again would produce somewhat stronger validity coefficient estimates) it is the step score that classroom teachers most commonly used.

### **Concurrent Validity**

We found moderately strong correlations between recorded spring step levels and end-of-year ITBS scores in reading at both second (0.52) and third grade (0.60). The relationship was slightly stronger (0.66) between step levels and students' reading performance on ISAT reading at the end of third grade.

### **Predictive Validity**

We also examined the predictive validity of the STEP scores from spring of second grade to student performance a year later at the end of third grade on both the ITBS and ISAT reading assessments. The end-of-second grade step levels correlated 0.58 with students' third grade ITBS scale scores in reading and 0.68 with their third grade ISAT reading scale score.

### **Establishing STEP Benchmark Validity**

The STEP assessment includes a set of developmental benchmarks (fall, winter, and spring) for each grade level covered by the STEP assessment system. These benchmarks are intended to guide teachers as to the adequacy of students' progress in learning to read over the course of their primary schooling. We routinely advise schools that any student who falls more than one step below their corresponding developmental benchmark should be considered for some form of supplemental services in order to accelerate their progress and increase the likelihood of achieving standards by the end of grade 3.

To examine the validity of this benchmarking system, we compared students' end-of-grade step levels in the spring of grades 2 and 3 to whether they achieved "at or above national norms" in reading on the ITBS (2<sup>nd</sup> and 3<sup>rd</sup> grade respectively) and whether they were categorized as "meeting or exceeding standards" on ISAT reading (3<sup>rd</sup> grade only.) The results are presented in Table 10 for the ITBS and in Table 11 for the ISAT. Each table highlights student performance relative to the benchmarks established by STEP for adequate developmental progress at that time of year and grade level. Specifically, for the spring of second grade, Step 9 is the benchmark and for the end of grade 3, Step 12 constitutes the developmental target.

**Table 10**  
**Percentage of Students who scored at or above the 50<sup>th</sup> percentile on ITBS by end-of-year step level**

Step level	6	7	8	9	10	11	12
2 <sup>nd</sup> Grade (n=167)	27%	37%	41%	82%	100%	100%	100%
3 <sup>rd</sup> Grade (n=123)	0%	0%	12%	17%	15%	54%	83%

**Table 11**  
**Percentage of Students Meeting or Exceeding Standards on ISAT by end-of-year step level**

Step level	6	7	8	9	10	11	12
3 <sup>rd</sup> Grade (n=123)	0%	0%	12%	15%	25%	50%	86%

Over 80 percent of the students who achieved Step 9 by the end of grade 2 had ITBS reading scores that were at or above second grade level. All 100 percent of the students (who were at Step 10 or higher on STEP) scored at or above grade level on the ITBS. In contrast, less than half of the students at Step 8 (41 percent) scored at or above

national norms on second grade ITBS reading. Correspondingly, over 80 percent of the student who achieved the benchmark of Step 12 for the end of third grade scored at or above national norms on the third grade ITBS. In contrast, only 54 percent of the students at Step 11 and 15 percent at step 10 met the same ITBS criterion.

Similarly strong results occurred in our comparisons of STEP benchmarks to student results on the ISAT reading. Over 85 percent of the students who achieved Step 12 by the end of third grade were categorized as meeting or exceeding standards on the ISAT reading assessment. In contrast, 50 percent of the students at Step 11 and only 25 percent at Step 10 met or exceeded state standards.

Taken together these data provide strong support for the use of STEP's developmental benchmarking system to guide programmatic decision making by teachers, principals and school district officials. Achieving STEP benchmarks are highly predictive of students' success on external standardized assessments in reading.

**Contact information:**

David Kerbow at University of Chicago: [d-kerbow@uchicago.edu](mailto:d-kerbow@uchicago.edu)

Anthony Bryk at Stanford University: [abryk@stanford.edu](mailto:abryk@stanford.edu)

Website: [www.iisrd.org](http://www.iisrd.org)

## References

- Adams, M. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Anthony, J. L., & Lonigan, C. J.. (2004). The nature of phonological awareness: Converging evidence from four studies of preschool and early grade school children. *Journal of Educational Psychology* 96:43-55.
- Bear, D. (1991). "Learning to fasten the seat of my union suit without looking around": The synchrony of literacy development. *Theory Into Practice* 30:149-157.
- Bear, D. R., Invernizzi, M., Templeton, S., & Johnston, F. (2000). *Words their way: Word study for phonics, vocabulary, and spelling instruction* (2<sup>nd</sup> ed.). Upper Saddle River, N.J.: Merrill.
- Bear, D., & Templeton, S. (1998). Explorations in developmental spelling: Foundations for learning and teaching phonics, spelling, and vocabulary. *The Reading Teacher* 52:222-242.
- Carver, R. P. (1990). *Reading rate: A review of research and theory*. New York: Academic Press.
- Carver, R. P., & Leibert, R. E. (1995). The effect of reading library books at different levels of difficulty upon gain in reading ability. *Reading Research Quarterly* 30:26-48.
- Chall, J. S. (1983). *Stages of reading development*. New York: McGraw-Hill.
- Clay, M. (1996). *An observation survey: Of early literacy achievement*. Portsmouth, NH: Heinemann.
- Clay, M. 1991. *Becoming literate: The construction of inner control*. Portsmouth, NH: Heinemann.
- Clay, M. 2001. *Change over time in children's literacy development*. Portsmouth, NH: Heinemann.
- Hoffman, J. V., Roser, N. L., Salas, R. Patterson, E., & Pennington, J. (2000). Text leveling and little books in first-grade reading. CIERA Report #1-010. Center for the Improvement of Early Reading Achievement, University of Michigan.
- Dole, J. A., Duffy, G. G., Roehler, L., Pearson, P. D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research* 61:239-64.

*DRP handbook: G & H test forms.* (1995). Brewster, N.Y.: Touchstone Applied Science Associates.

Ehri, L. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading* 18: 116-25.

Ehri, L. (1997). Learning to read and learning to spell are one and the same, almost. Pp. 237-294 in *Learning to spell: Research, theory, and practice across languages*, C. Perfetti, L. Rieben, M. Fayol, eds. Mahwah, N.J.: Lawrence Erlbaum Associates.

Fountas, I. C. and Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann.

Fountas, I. C. and Pinnell, G. S. (2001). *Guiding readers and writers grades 3-6: Teaching comprehension, genre, and content literacy*. Portsmouth, NH: Heinemann.

Frith, U. (1985). Beneath the surface of developmental dyslexia. Pp. 301-330 in *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*, K. Patterson, J. Marshall, and M. Coltheart, eds. London: Lawrence Erlbaum Associates.

Ganske, K. (2000). *Word journeys: assessment-guided phonics, spelling and vocabulary instruction*. New York: Guilford Press.

Henderson, E. (1990). *Teaching spelling* (2<sup>nd</sup> ed.) Boston: Houghton Mifflin.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Bray, G. B., Naylor, R. J., Lewis, J. C., Ordman, V. L., & Qualls, A. L. (2001). *The Iowa tests Interpretive guide for teachers and counselors*. Itasca, IL: Riverside Publishing.

Illinois State Board of Education, Division of Assessment. (2004). *The Illinois State Assessment 2004 Technical Manual*. Springfield, IL: Author.

Invernizzi, M., Abouzeid, M., & Gill, T. (1994). Using students' invented spelling as a guide for spelling instruction that emphasizes word study. *Elementary School Journal* 95:155-67.

Johnson, M. S., Kress, R. A., & Pikulski, J. J. (1987). *Informal reading inventories* (2<sup>nd</sup> ed.). Newark, DE: International Reading Association.

Kerbow, D. (1999) A school-based literacy assessment system: The technical core of progress for all. Paper presented at the American Education Research Association, Montreal, Canada.

Kerbow, D., Gywne, J., & Jacob, B. (1999). Evaluation of literacy achievement gains at the primary level. Paper presented at the American Education Research Association, Montreal, Canada.

- Koslin, B. L., Zeno, S., Koslin, S. (1987). *The DRP: An Effectiveness Measure in Reading*. Brewster, N.Y.: Touchstone Applied Science Associates.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology* 6:293-323.
- Lundberg, I., Frost, J., & Peterson, O-P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly* 23: 264-84.
- McBride-Chang, C. (1998). The development of invented spelling. *Early Education & Development* 9:147-60.
- McBride-Chang, C. (1999). The ABCs of the ABCs: The development of letter name and letter sound knowledge. *Merrill-Palmer Quarterly* 45:285-308.
- Morris, D. (1993). The relationship between children's concept of word in text and phoneme awareness in learning to read: A longitudinal study. *Research in the Teaching of English* 27: 133-57.
- Morris, D., Bloodgood, J., Jomax, R., & Perney, J. (2002). Developmental steps in learning to read: A longitudinal study in kindergarten and first grade. *Reading Research Quarterly* 38:302-28.
- Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, R. B., & Beatty, A. S. (1995). Listening to Children Read aloud: Data form NAWP's Integrated Reading Performance record 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.
- Pressley, M. (1998). *Reading instruction that works: The case for balanced teaching*. New York: Guilford Press.
- Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99-136). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Snow, C., Burns, M., & Griffin, P. (Eds.). (1998) *Preventing reading difficulties in young children*. Washington, D.C.: National Academy Press.
- Stenner, A. J. (1996). Measuring reading comprehension with the lexile framework. Paper presented at the North American Conference on Adolescent/Adult Literacy, Washington, DC.
- Torgesen, J. K., & Davis, C. (1996). Individual difference variables that predict responses to training in phonological awareness. *Journal of Experimental Child Psychology* 63:1-21.

Treiman, R. ; Tincoff, R., Rodriguez, K., Mouzaki, A., & Francis, D. J. (1998). The foundations of literacy: Learning the sounds of letters. *Child Development* 69:1524-540.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading. *Psychological Bulletin* 101:192-212.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development* 69:848-72.

Wright, B. D., & Master, G. N. (1982) *Rating scale analysis*. Chicago: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.