Section 5

Technical Features of the Critical Reading Inventory: Development and Validation

Construction of the Components of the CRI

The primary purpose of the CRI is to assist teachers and reading specialists in their assessment of children's reading ability, strengths, needs, and interests. To determine the extent to which this purpose is met, we conducted extensive field testing of the original test materials, including word lists; interviews for children, parents/guardians, and teachers; narrative and informational text selections; three types of comprehension questions; and retelling rubrics and a fluency rubric.

Test materials were evaluated by 20 practicing reading specialists, who averaged 15 years of experience and who administered informal reading inventories an average of 100 times per year. All had at least one master's degree and two were enrolled in doctoral programs. All were certified reading specialists, with 14 also certified in elementary education, 6 in secondary education, 4 in special education, 2 in early childhood education, and 7 seeking educational leadership certification. Their teaching assignments included urban, suburban, and rural districts ranging in grade level from Kindergarten to Grade 12, ranging in family income from low to high, and including several schools with diverse populations.

In addition to this wide-ranging professional evaluation, the authors engaged in an extensive analysis of student responses to comprehension questions and retellings. Our objective was twofold: (1) to identify items and passage segments that were not effective in discriminating between accomplished and struggling readers, and (2) to identify patterns of thinking that could guide users in the instruction of all readers (Applegate et al., 2006).

As a consequence of the input from the initial field testing and the analysis of responses, we made numerous changes in passage wording, readjusted several reading levels, increased the clarity of directions for administration and scoring, and revised numerous comprehension items for both narrative and informational assessment.

Word Lists

The Word Lists in the CRI were designed to allow users to gain insights into two dimensions of a reader's context-free word recognition ability. The first 10 words in each graded word list were selected at random from high-frequency word lists such as Fry's Instant Words, the Dolch word lists (Dolch, 1948), Frances and Kucera's word lists (1982), and the core reading vocabulary lists compiled by Taylor et al. (1989). The final 10 words in each list were drawn from the graded narrative passages. This format is designed to allow users to compare the reader's word recognition in isolation to word recognition within the context of reading. We used only narrative passages as sources for these words in order to avoid very specific content-related vocabulary that may underestimate the starting grade level for reading. Words that appear in the passages are marked with an asterisk in the Examiner's Copy of the Word Lists.

CRI 2 includes two sets of word lists, Form A and Form B, that allow users to administer the word lists in pretest and posttest format.

Interviews

We created the Student Interview questions as a means of gathering (1) background information on the child, (2) information about the child's view of reading, and (3) information about the child's interests and attitudes. Our intent was to add a more balanced and affective dimension to the analysis of the data collected. To enable users to obtain multiple perspectives rather than just a self-report, we included both a Parent/Guardian and a Teacher interview form. The Teacher Interview is designed to uncover the teacher's perception of the child's ability and classroom performance, a description of the classroom reading program, and an indication of the teacher's theoretical orientation to reading instruction. Parent/Guardian information includes the parent or guardian's perception of the child's reading and classroom performance, the level of parental involvement, and literacy/school activities in the home. These sources contribute to a triangulation of key information to lend to the CRI 2's credibility and practical value in diagnosis. Examples of typical responses to interviews are included in the case studies on the website that accompanies the CRI.

Graded Passages

All graded passages were written by the authors based on 5 years of observational notes in classrooms and clinics for Grades K–12 and over 20 years of professional experience in classrooms per author, including supervision of reading practicum experiences at their respective institutions. These observations included notes regarding curriculum, materials, topics, themes, and children's interests and developmental responses toward the curriculum and activities, in addition to personal experience. Key ideas for the content of the passages were brainstormed among the authors; then each author wrote several passages on agreed-upon themes or content for a particular age or grade level. We then met, shared, and revised the passages collaboratively. Three narrative and three informational passages were written for each level, Pre-Primer through Grade 12. At the lower levels, Pre-Primer through Grade 1, we made an effort to include repetitive text and structural elements that would be appropriate for the emergent reader (Fountas & Pinnell, 1996). Throughout the CRI 2, but particularly at the higher level of the informational passages, we attempted to include high-interest but less familiar topics in order to avoid potential reader overexposure to topics, a difficulty we had noted in other recently published IRIs.

Passage Readability

We initially evaluated passage readability using multiple formulas included in the MicroLight software for readability, but we found many inconsistencies across formulas. We decided to use the Flesch-Kincaid Formula found on Microsoft Word software (1997). The formula itself is similar in formulation to the Fry Readability Scale and the SMOG Readability Formula (*http://www.med.utah.edu/pated/authors/readability/html*). This choice allows the CRI users easy access for checking the readability of materials they wish to use with their students. It is important to note that the 1997 software package was used to determine CRI passage readability. Subsequent versions of the Flesch-Kincaid Formula on later editions of Microsoft Word tend to significantly overestimate readability. Exact information regarding readability and passage length is provided on the introductory page of the Examiner's Copy of both the narrative and informational passages. Passages were written in an attempt to maintain a .2 or less difference between the readability estimates at each level and to ensure that there would be equal intervals in readability between levels (Gerke, 1980; Klesius & Homan, 1985).

In addition to readability considerations, we used content, appeal, interest, background knowledge, text structure, overall length, and print format/layout in determining appropriateness for each passage at each grade level. It is well established that sentence length, word frequency, and number of syllables and words per sentence are not the only factors that should be considered in evaluating the difficulty of text (Jitendra et al., 2001; Johnston, 1997; Kinder, Bursuck, & Epstein, 1992; Lynch, 1988).

Pictures were included at the Pre-Primer through Grade 1 levels for several selections of narrative passages, and photographs were included for some passages at the same levels for the informational passages. Thus, the examiner can compare a child's performance with and without picture clues at each of these three levels. The pictures and photographs were specifically created to support the text in a meaningful way.

For the narrative passages, we used a variety of genre types, including familiar childhood experiences as well as folk tales and fablelike stories. All passages were created by the authors in order to eliminate the concern for past exposure to stories by our intended audience. Passages were written with an eye toward gender fairness and ethnic diversity. The informational texts include original social science and science content as well as biography. All informational selections were thoroughly researched for accuracy of the facts presented.

During field testing, we also assessed children's interests, background knowledge, and performance on the materials as a means of determining level appropriateness. As we discussed earlier, several passages were modified for ease or difficulty.

Selection length was also an issue that we discussed and examined. Based on previous recommendations (Jitendra et al., 2001; Johnson et al., 1987; Jongsma & Jongsma, 1981; Klesius & Homan, 1985), it is important to have adequate length (more than 125 words) at each level except for Pre-Primer so that there will be substantive content to serve as the basis for comprehension assessment. In addition, passage length is important in that the majority of materials that children will be expected to read in their classrooms (from mid-first grade on) will have well over 100 words. State and national assessments also tend to include longer passages. Research has suggested that shorter IRI selections tend to overestimate children's reading ability for regular classroom materials and show inaccurate miscue patterns (Bowden, cited in Klesius & Homan, 1985). Jitendra et al. also found that most recent basal series have selections that are well over 100 words in length and most content text chapters also fit this pattern. Because the intention of IRIs is to estimate placement into curriculum materials, suitable length was an important factor in the development of the CRI.

Assessment of Comprehension

Because comprehension is the most important aspect of reading that is being evaluated, we developed two tools to assess comprehension: retelling rubrics and postreading questions. An Oral Reading Fluency Rubric is included to accompany assessment of the reader's comprehension.

Retelling Rubrics

Rubrics for the scoring of retellings are included for each passage in the CRI. In the case of narrative selections, the elements assessed in the retelling are based on essential story grammar components (i.e., characters, character goals or problems, and the key steps characters take toward the solution of the problem or the attainment of the goal). The final component of the retelling rubric is the presence or absence of a well-supported personal response to the text.

Rubrics for informational passages draw a distinction between macro-concepts and micro-concepts. Macro-concepts are defined as superordinate ideas that are central to the information presented in the text and which are supported, illustrated, or further explained by additional information in the text. The supporting details are defined as micro-concepts, not because they are unimportant but because they act as support for the more general and inclusive macro-concepts.

In the case of either narrative or informational text, test users will consult the retelling rubric that accompanies each passage. Users must note the presence or absence of each of the text elements listed in the rubric, and then consult the Retelling Scoring Guides in the Examiner's Tools section of the CRI Manual. These guides provide detailed instructions for assigning a numerical score to a retelling. Users also have the option of downloading the Automated Scoring and Interpretation Interview (ASII) from the CRI website. The ASII automatically calculates retelling scores.

The formulas included in the guides were developed as a part of the authors' analysis of student responses to retellings and comprehension items. Our objective was to create a reliable and valid numerical representation of a retelling score that could inform users at a glance of a reader's proficiency in unaided recall of text. We will discuss the results of our tryout of the scoring guides in our subsequent discussion of the standardization study of the CRI.

Because reader response to text is central to our definition of critical reading, we have included an invitation to readers to discuss their personal response to the text. Assessment of the personal response is included in the rubric used for scoring retellings. During field testing of the CRI, this component was identified by participants as a positive feature that provided reliable information about a reader's view of and attitude toward reading.

Postreading Questions

After examining several current IRIs and analyzing their methods of comprehension assessment, we determined there was a need for a more structured assessment higher-level thinking in the form of inferential and critical response items (Applegate et al., 2002; Manzo, Manzo, & McKenna, 1995; Nessel, 1987). We then applied the questioning criteria we had developed (see Figure 5–1), along with the recommendations from Klesius and Homan (1985), to create Text-Based, Inference, and Critical Response questions for each passage. All questions are labeled as to type and possible correct responses, many of which were actually given by children during the field testing. These are included on the Examiner's Copy of the CRI.

For levels Pre-Primer through Grade 1, there are 8 questions for each passage at each level and for each type of text. For Grades 2 through 12, there are 10 questions for each passage at each level for each type of text. For narrative passages, when there are 8 questions, there are 3 Text-Based, 3 Inference, and 2 Critical Response items; when there are 10 questions, there are 4 Text-Based, 3 Inference, and 3 Critical Response items. For informational text, when there are 8 questions, there are 3 Text-Based, 3 Inference, and 2 Critical Response items; when there are 10 questions, there are 4 Text-Based, 4 Inference, and 2 Critical Response items. There are more Critical Response items for narrative text at the upper levels because there is more opportunity for this type of thinking in response to this type of text (Rosenblatt, 1983). No direct assessment of vocabulary was done because these types of questions are often dependent on a child's background knowledge and language ability (Duffelmeyer, Robinson, & Squier, 1989).

Virtually all questions require that the child actually read the passage and use information from it in order to avoid one of the pitfalls of many IRIs regarding passage dependency (Klesius & Homan, 1985). No questions rely totally on a child's past experience or background knowledge, although we know that some children will and should make use of their experiences in order to help them comprehend and respond to the passages. Passage dependency of questions was also evaluated by asking 16 children and 6 adults the comprehension questions without affording them the opportunity to read the passages. More than 90% of the questions could not be answered. Questions from the informational passages at the lower levels (Pre-Primer through Grade 2) were more readily answered correctly without reading the passages; however, this was done more frequently by older children and adults who for the most part would not be evaluated by those passages.

We also noted that lookbacks could be utilized as an additional dimension of the assessment of reading comprehension. Lookbacks can be most useful with students who have recurring difficulty with literal recall and who frequently respond with "I don't know" or "I don't remember." Lookbacks enable the user to assess whether the reader is experiencing difficulty with remembering or with comprehending the text. It also enables you to assess the reader's ability to quickly locate information relevant to questions that he or she had been unable to answer (Alvermann, 1988; Bossert & Schwantes, 1995–1996; Swanson & De La Paz, 1998).

Estimating Reading Levels

We have chosen to use the Betts criteria for estimating reading levels, even though there are studies that have both supported (Bader & Wiesendanger, 1989; Johns, 1991; McKenna, 1983) and critically examined them (Johns & Magliari, 1989; Lowell, 1969; Pikulski, 1990). Because they

Text-Based Question Types (TB):

1. **Literal Items:** Answers to these items are stated explicitly (verbatim) in the text. They simply require that the readers recall what they have read.
2. **Low-Level Inference Items:** The answers to low-level inferences are not stated verbatim in the text but may be so close to literal as to be very obvious. All inference items require that readers draw a conclusion on the basis of the text and use their background experiences to some extent as well. However, low-level inferences require very little in the way of drawing conclusions. We classified as low-level inferences, for example, items that
    - involve the recognition of information in different words from those used in the original text. Such items require of the reader only a translation of the printed text.
    - require the reader to identify relationships that exist between ideas in the text. Such items as these are not literal only because the writer has not made the relationship explicit by using a grammatical marker (e.g., *because*). This is not to say that the skill of making such connections is unimportant. Classification of an item as low-level merely reflects that the writer assumes that at a given grade level, the reader can and will make the connection.
    - deal with details largely irrelevant to the central message of the text.
    - require that the reader draw solely on background knowledge or to speculate about the actions of character without the benefit of information in the text that may transform speculation into a logical prediction.

Inference Question Types (I):

3. **High-Level Inference Items:** These items call for the reader to link experience with the text and to draw a logical conclusion. Answers to these items require significantly more complex thinking than low-level inferences. Examples include those items that require the reader to
    - devise an alternative solution to a specific problem described in the text.
    - describe a plausible motivation that explains a character's actions.
    - provide a plausible explanation for a situation, problem, or action.
    - predict a past or future action based on characteristics or qualities developed in the text.
    - describe a character or action based on the events in a story.

Critical Response Question Types (CR):

4. **Response Items:** These items call for a reader to express and defend an idea related to the actions of characters or the outcome of events. Response items differ from high-level inference items in that they are usually directed toward broader ideas or underlying themes that relate to the significance of the passage. Although high-level inference items are directed toward a specific element or problem in the passage, response items require a reader to discuss and react to the underlying meaning of the passage as a whole. Examples include items that ask the reader to:
    - describe the lesson(s) a character may have learned from experience.
    - judge the efficacy of the actions or decisions of a character and defend the judgment.
    - devise and defend alternative solutions to a complex problem described in a story.
    - respond positively or negatively to a character based on a logical assessment of the actions or traits of that character.

Figure 5–1

Criteria for Determining Question Types

are the most frequently used criteria, we believe that they are useful for comparative purposes. In addition, many of the other criteria that have been tried are very close to Betts's original criteria. We wish to emphasize that, in any case, these criteria are only guidelines for estimating levels. Many other factors enter into such estimates, including qualitative information, knowledge about the student's actual performance in the classroom, and observations about reading strategy use and behavior (Goodman, 1997; Harris & Lalik, 1987).

Reliability and Validity of the CRI: Standardization Study

We assessed several dimensions of the reliability and validity of the CRI by conducting a broad standardization study that involved the administration of the CRI to 215 students ranging from Grade 1 through Grade 12. The following analyses were carried out on 1,255 passages administered to this standardization sample. The sample for this study included a nearly equal proportion of males and females, representing a wide range of ethnic groups, most of whom attend public schools in the tristate area including Pennsylvania, New Jersey, and Delaware. The sample included just over one quarter of its students who had been identified

Table 5–1 Demographic Characteristics of Standardization Sample (N = 215)

Gender	Male	Female			
	$N = 105$	$N = 110$			
Ethnicity	Caucasian	Black	Hispanic	Asian	Other
	$N = 150$	$N = 38$	$N = 15$	$N = 6$	$N = 6$
School	Public	Private	Parochial		
	$N = 156$	$N = 21$	$N = 38$		
Reading level	High	Average	Low	N/A	
	$N = 56$	$N = 68$	$N = 89$	$N = 2$	
Grade level	1 to 3	4 to 8	9 to 12		
	$N = 93$	$N = 95$	$N = 27$		

by teachers or parents as high-achieving readers. The remaining three quarters of the sample fell into the average and low-achieving groups, those most likely to be assessed by an informal reading inventory. Thirty of the 215 students in the sample were eligible for special education services, whereas 5 were enrolled in programs for the gifted. For 14 of the students in the sample, English was not their native language. Demographic data for the sample are included in Table 5–1.

All tests were administered by graduate students who had been trained in the use of the CRI through demonstrations of test administration similar to those found on the DVD that accompanies the main text. In addition, all students had completed several tutorials for scoring miscues, retellings, and comprehension items, tutorials that are included on the same DVD.

Interrater reliability for scoring the CRI was assessed by comparing the scoring of miscues, retellings, and comprehension items of the test administrators to those of six trained and experienced experts in the administration and scoring of the CRI. Three of the expert scorers were practicing reading specialists; three were former reading specialists. All expert scorers worked independently. The aim of the comparative scoring study was to establish the extent to which users of the CRI can, with a few hours of instruction and practice, administer the instrument correctly and score the results with consistency.

Word Lists

Whereas the ability to administer and score the Word Lists segment of the CRI is a straightforward matter, there still remains the issue of whether the lists achieve what they are designed to do. We focused on two specific issues in our validity study of the lists. First of all, users need assurance that the lists become progressively more difficult as the grade levels increase. Second, because we use the lists specifically to determine a starting point for the reading portion of the CRI, we needed to determine if they can act as accurate indicators of that starting point.

To address the first question, whether the lists become increasingly difficult, we considered presenting data drawn from the graded word lists that we used as sources to demonstrate increasing difficulty of the words on our lists. However, the standardization study data provide more accurate and direct evidence of the issue. We reasoned that if the word lists are progressively more difficult, that fact would be reflected in the scores that students achieved during the flash and untimed presentation of the lists. Specifically, we assumed that the progressive difficulty of the word lists would be demonstrated if children who read consecutive graded word lists would achieve progressively lower scores as the grade level increased.

We examined the Flash and Untimed Word List scores of all 215 subjects and compared the scores of lists administered at consecutive grade levels. We then simply tallied the number of times that a score on a higher grade level was matched or exceeded by the score at the lower grade level. For example, if a reader was given a third-grade word list followed by a fourth-grade list, we would not expect the fourth-grade score to exceed the score on the third-grade list that was designed to be easier.

An analysis of the data from the standardization study yielded a total of 864 paired comparisons. In 804 cases, or 93.1% of the total observations, the score from the higher-grade-level list did not exceed the score from the lower-level list. We concluded from this analysis we could have a high level of confidence in the progressive difficulty of the CRI Word Lists.

What remained to be demonstrated was the usefulness of the lists in identifying a starting point for the onset of the CRI oral reading and reading comprehension measures. We advise CRI users to begin the reading test at the highest level where the reader achieves a perfect score of 100% during the flash presentation of the words. Our reasoning is that if readers can flawlessly identify a list of words drawn from instructional materials at a given grade level, the chances are good that they will be able to read those materials with a high degree of competence. Of course, as any teacher knows, flawless word recognition does not guarantee flawless comprehension. However, the investment of time involved in testing a child's reading and comprehension makes any good estimate of a starting point worthwhile.

We examined the standardization study data and found that 200 subjects had achieved a perfect score on the flash presentation of the word lists. When we assessed their reading performance at that beginning level, we found that 84.5% of the students were reading at their instructional or independent level. Thus we concluded that the use of the Word Lists as an estimate of a beginning level for the CRI, although by no means perfect, was both valuable and valid.

Scoring of Miscues

Reliability

A total of 604 of the passages administered to the standardization sample were administered as oral reading. The examiners were instructed to use a tape recorder to capture the oral reading of the children and to note any deviation from the text on the Examiner's Copy of the passage in question. The examiners were then instructed to distinguish between miscues that maintained meaning in the context of the passage (to be marked with a plus sign) and miscues that violated meaning (to be marked with a zero). The job of the expert scorer was to determine if he or she agreed with the judgment of the graduate students who administered the tests and scored the miscues.

A total of 3,827 miscues were noted in the oral reading of the subjects and the expert scorers agreed on the assessment of the examiners 3,624 times for a total interrater reliability percentage of 94.7. This very high level of reliability can be explained in part by the nature of the task. We did not ask graduate students to identify the source of miscues or to analyze miscues at any deeper level other than to judge whether they maintained or violated meaning. In the CRI we interpret the tendency of readers to violate meaning in their oral reading as indicative of a potential loss of comprehension. Because the percentage of meaning-maintaining miscues is the primary component of the Meaning Maintenance Index (MMI), these results suggest that teachers can, without inordinate amounts of instruction, identify meaning-violating and meaning-maintaining miscues with a high degree of reliability.

Validity

The Reading Accuracy Index (RAI) is equivalent to a form of reading assessment that has long been associated with the use of informal reading inventories to estimate a child's reading grade level (Betts, 1954; Johnson et al., 1987). The RAI is a simple baseline measure of the percentage of words read aloud that accurately replicate the text. The MMI, on the other hand, allows the examiner to distinguish between types of miscues and to assess the percentage of words read aloud that represent an attempt on the part of the reader to maintain the sense of the language. Thus, the MMI would seem on the surface to be a more useful tool in level setting because a high MMI suggests a reader who views the task of reading as one of constructing a meaningful message in response to text.

The problem with using the MMI for level setting is that there are virtually no data to support its use or even to validate its worth as an educational construct. We designed the data analysis that follows as an attempt to establish the MMI as a worthwhile estimate of a reader's view of reading as a meaningful activity. We reasoned that if two readers made equal numbers of miscues, but Reader A's miscues preserved the syntax and sense of language whereas Reader B's frequently violated them, then there would be some measurable difference in the overall comprehension of these two readers. Otherwise, there would seem to be little value in distinguishing between the RAI and MMI.

We reviewed the CRI protocols that had been administered by our graduate reading specialist candidates for the past 2 years. The sample of students tested represented an exceptionally wide range of students and an equally wide range of socioeconomic and achievement levels in reading and language arts. Our students selected subjects from among their own students, the children of friends and family, and students in schools in which they happened to be working.

We posited the following distinction between levels of achievement in the RAI and MMI: If a pair of readers read the same passage and achieved the same RAI level, it would mean that they had made nearly the same number of miscues. If, however, one reader's MMI score was at least two percentage points lower that the other's, it would mean that the miscues made by the latter reader included a significant number of violations of sense. For example, if a pair of readers read a passage that was 300 words in length, and achieved an RAI score of 97, that would mean that each of them had made 9 miscues. If the first reader scored an MMI of 100, it would mean that virtually none of his miscues violated sense. If the second reader scored an MMI of 98, it would mean that 6 of her miscues represented a distortion of semantics. We reasoned that such a number of meaningful distortions would be likely to affect the reader's comprehension of the text.

We asked a graduate assistant to review the records of the administration of the CRI to more than 400 students, ranging from Grade 1 through Grade 12. We asked her to seek out matched pairs of students who had read the same passage aloud, had the same RAI score, but had an MMI score that differed by no fewer than two percentage points. We did not reveal to her the reason for her search or the details of the analysis that we had planned. Thus it was not likely that she intuited the purpose of the search or could use that intuition to screen the comprehension scores of potential candidates for the study.

The screening yielded 32 matched pairs of readers, with the grade level of the oral passages they read ranging from Pre-Primer to Grade 6. The retellings and responses to comprehension questions were graded by two expert scorers working independently. The level of interrater agreement was 96% for the comprehension items and 95% for the retellings with all differences resolved by discussion.

The results of the analysis were unequivocal. For the high MMI group, 28 of the 32 readers scored higher in their overall percentage of comprehension items scored correctly, with an average difference in scores of 28.94%. The results of comparative analyses are presented in Table 5–2. A one-way Analysis of Variance (ANOVA) suggested that differences between groups were highly significant ($F = 61.21$, $p < .0001$). We must note that a comparison of comprehension scores based on only 8 to 10 comprehension items and a single retelling is not an optimum situation. However, the powerful and consistent results seemed to us to minimize this disadvantage. This dramatic difference in comprehension scores seems to confirm what many reading professionals may view as the obvious: Readers who view reading as a meaning-making process are likely to be more successful at comprehending materials than readers who do not demonstrate that view in their processing of text.

A comparison of the retelling scores of both groups yields equally compelling, if not equally dramatic evidence. For the high MMI group, 26 of the 32 readers scored higher in their retelling scores with an average difference in scores of more than 1.25 points on a scale of 4.00. A one-way ANOVA suggested that differences between groups were highly significant ($F = 52.30$, $p < .0001$).

Table 5–2 Comparison of Reading Comprehension and Retelling Scores for High MMI and Low MMI Readers

Mean Score: Comprehension

High MMI (*N* = 32)	Low MMI (*N* = 32)	Difference	*p*Ratio	Significance
77.5%	48.6%	28.7%		

Mean Score: Retelling

High MMI (*N* = 32)	Low MMI (*N* = 32)	Difference	*p*Ratio	Significance
2.02	.81	1.21		

In retrospect, we were not surprised by the results of our analysis. It stands to reason that text comprehension is clearly related to the extent to which a reader views reading as requiring comprehension. The high MMI readers avoided the types of miscues that violated sense precisely because they were monitoring their reading to ensure that it made sense. These preliminary results provide some encouraging evidence for the overall worth of the construct of the MMI, and suggest that further research may be warranted into its value as a factor in the estimation of reading levels.

Scoring of Comprehension Items

Reliability

A major issue in the standardization study of the CRI was the extent to which expert scorers would agree with the examiner scoring of comprehension item responses. This is an issue that cuts to the heart of the CRI as a useful assessment tool because the proportion of higher-order items in the CRI significantly exceeds that of other informal reading inventories (Applegate et al., 2002). By definition, text-based items and low-level inferences have a clearly stated or obviously implied answer, but inference and critical response items in the CRI lend themselves to far more creative and challenging responses, and often allow for multiple correct and partially correct responses. The ability of examiners with a reasonable level of professional preparation to score these items with reliability can go a long way toward determining the value and usability of the instrument as a whole.

Examiners were prepared to score comprehension items largely via the completion of 12 tutorials very similar to those included on the site that accompanies the CRI. Examiners were instructed to score each response as meriting full credit, partial credit (1/2), or no credit. Expert scorers then independently assessed each response and determined whether they fully agreed, partially agreed, or completely disagreed with the scoring of the examiners. The interrater reliability was calculated as a percentage of agreement between expert and novice users.

The six expert scorers assessed a total of 11,905 responses and agreed on the scoring of 11,328 for a total percentage of agreement of 95.2. When results were divided between narrative and informational passages, the total percentages of agreement were 94.8 for narrative passages and 96.1 for informational passages. These are exceptionally high levels of agreement and suggest that the CRI can be used with confidence by professionals who have a modest but reasonable level of preparation in the interpretation and scoring of item responses. These data also suggest that the tutorials that accompany the CRI are useful tools in preparing professionals for the reliable scoring of comprehension items.

Validity: The Existence of Different Item Types

Davis (1968), in his landmark attempt to isolate distinct factors in the assessment of reading comprehension, discussed at some length the nature of reading as a unitary process. Specifically, he suggested that further attempts to empirically validate different types of comprehension items may be an exercise in futility for several reasons. First of all, there is powerful evidence that reading develops as a unitary skill, with the vast majority of children learning to read in an integrated fashion, combining skill development in word recognition with the development of a full range of comprehension skills. More specifically, it seems that most children will seamlessly transition from an oral language system that is pragmatic and meaning-centered to a written language system that requires the same levels and types of thoughtful response.

If Davis is correct, then the attempt to isolate different types of thinking skills and empirically validate their existence in a broad sample of the population is indeed doomed to failure. For the overwhelming majority, reading assessment that requires any purportedly distinct levels of thinking is still rooted in an act that begins with the extraction and construction of meaning in response to text. Any differences evidenced among a relatively small number of children are sure to be washed out by the vast majority of children for whom thinking about what they read is as natural as thinking about their life experiences. Thus it may not be possible for mathematical analyses to establish once and for all the existence of different types of thinking in response to text. It may well be that the items themselves are highly interrelated; they simply elicit different responses from different groups of readers.

However, teachers in real-world classrooms must deal with real-world issues. Suffice it to say that it matters little to Jimmy and Sharon in my third-grade class that their inability to evidence any thinking about what they have read has not been empirically validated by mathematical analysis. They simply do not realize that the very type of thinking that they evidence in their daily interactions with others is that type of thinking called for in the reading that they do in school and home. They struggle with any circumstances that require more than a cursory reaction to text. Their response journals include nothing more than summaries of text, in contrast to the thoughtful responses of their classmates, and they often lapse into silence while other students are discussing their reactions to the ideas presented by the authors of the texts they have read.

The truth is that we have in our instructional repertoires techniques and activities that can help Jimmy and Sharon realize that they can and must apply their thinking skills to the reading that they do. One of the most short-sighted things that we can do as reading professionals is to demand of research and mathematical analysis answers that they simply are ill-equipped to supply. If we refuse to acknowledge the existence of different levels of thinking in reading, we will likely not require of our students anything more than a mastery of factual information. Indeed, if Allington (2001) is correct, the overwhelming majority of questions that teachers ask about what children have read require only memory for details. If that trend continues, we will never discover the needs of children like Jimmy and Sharon and we will never be able to help them grow in their ability to react and respond to text.

Scoring of Retellings

Reliability

Graduate students were instructed to audiotape all retellings and transcribe them verbatim to their examiner's materials. They were then instructed to use the Retelling Rubric that accompanies each passage in the CRI and to determine the extent to which passage elements in the rubric were present in the retelling. Each graduate student was given a copy of the Scoring Guides for Retellings found in the Tools section of the CRI Manual. Some students used the automated retelling help feature of the ASII to calculate retelling scores; others used the Scoring Guides.

Retelling Rubrics

Reliability

The retelling rubrics in the CRI 2 were developed to follow the structure of a common story grammar in the case of narrative text and a structure of main ideas (macro-concepts) and supporting detail (micro-concepts) in the case of informational text. Our study of authentic student retellings led us to make several modifications and culminated in the development of detailed Guides for Scoring Retellings. We set out to validate the use of our numerical retelling scores by comparing them to the total percentages correct in the comprehension items that accompany each passage. Our reasoning was that if the retelling scores were completely unrelated to the reader's comprehension, they would be of little value as a diagnostic tool.

We began our investigation with the notion that a retelling cannot and should not be used as the sole source of comprehension assessment for any student for several reasons. First and foremost among these is the fact that students experience widely varying levels of exposure to story structure specifically and to text structure as a whole. Many children, for example, are taught the elements of a story grammar and exposed to a full range of reading and writing activities in which they are expected to apply it. Other children will have virtually no idea of its existence. Needless to say, the narrative retellings of these two groups will vary widely in terms of overall quality and effectiveness. Still other groups of readers have learned to rely on the structure provided by a set of comprehension questions and have little or no idea of how to structure their own unaided recall of text. For these reasons we hypothesized that the correlation between the retelling scores calculated using the

CRI 2's scoring rubric and the comprehension item percentage scores would be statistically significant and moderate in size.

Our sample of 215 students read 1,255 passages and examiners recorded a retelling for each passage. The retelling scores assigned by the examiners were cross-checked by an expert scorer, using the automatic retelling scoring device included as part of the Automated Scoring and Interpretation Interview (ASII) to ensure consistency of scoring. In 92.5% of the passages, the test administrator and the expert agreed on the scoring of the retelling, establishing the CRI Retelling Rubric as an instrument that can be used reliably by professionals with a reasonable level of preparation in the use of the instrument.

Validity

To establish the validity of the Retelling Rubrics and Scoring Guides of the CRI, we correlated the retelling score with the total comprehension item percentage for each passage. In cases where disagreement existed between the graduate student and the expert scorer, the expert's score was used. An analysis of 905 narrative passages revealed a correlation coefficient of .51 ($p < .001$) between the retelling score and the total comprehension item score. An analysis of 352 informational passages revealed a correlation of .43 ($p < .001$).

These results were certainly in line with our expectations. There is clearly a logical relationship between a reader's unaided recall and ability to respond to comprehension questions. It is very unusual to see a reader fail miserably in responding to comprehension questions after she has delivered a competent and thorough retelling. And there certainly appears to be some measurable common variance that exists between retellings and performance on comprehension questions. However, variations among readers in their experience with retellings and in the type of retelling instruction they may have received makes the use of retellings suspect when they are used as the sole measure of a reader's comprehension. Nevertheless, the common variance of unaided and aided recall lends value to the numerical expression of the quality of retelling assessed by the CRI.

The difference that exists in the correlations between retellings and comprehension items for narrative and informational text is sizeable enough to suggest some variance that is not accounted for by the sheer similarity of the tasks. Informational texts can be organized in several different ways (e.g., cause-effect, examples, comparison, contrast, etc.) and even in combinations of structures. Stories, on the other hand, are characterized by a fairly universal structure often referred to as a story grammar. Some experts have suggested that the ability to apprehend the story structure is common even among young children. It stands to reason that children would be more aware of and comfortable with the narrative as opposed to the informational structure.

This study did not include enough cases to warrant the examination of the relationship between retellings and comprehension scores at various grade levels. Some researchers have reported a shift in the correlation from lower values in the lower grades to higher values in the upper grades (Leslie & Caldwell, 2006).

Fluency Rubric

Reliability

The Fluency Rubric included in the CRI was developed as a consequence of extended analysis of the oral reading of students representing a wide cross-section of grades and ability levels. We set out to develop an instrument that could help teachers distinguish clearly among several different proficiency levels in reading fluency. To be a useful tool, a fluency rubric must enable users to assign a value to a child's oral reading with consistency and clarity. The descriptions included in the rubric must be sufficiently clear to allow users to identify characteristics of a child's oral reading and match them accurately with the weighted descriptions. To test the reliability of the CRI rubric, we randomly selected 30 audiotaped oral readings from among the subjects in the standardization study and compared scores assigned by graduate students to those calculated by expert scorers.

The graduate students who participated in the standardization study were instructed to record the oral reading, retellings, and responses to comprehension questions of the students that they tested. After they had analyzed and scored the results of the CRI, they estimated reading levels, using the criteria described in the CRI Manual. They were then instructed to identify the reader's highest instructional grade level, retrieve the tape of the child's oral reading at that grade level, and match the rubric descriptions to the characteristics of the child's oral reading. They then totaled the point values associated with the descriptions to arrive at a final fluency score.

The 30 audiotaped readings selected for this study were distributed to three experienced and expert CRI users. Each was provided with a printed copy of the targeted text and a copy of the CRI Fluency Rubric. The experts were asked to assess the oral reading and assign appropriate numerical values to the retelling without knowledge of the score originally assigned by the graduate student. We believed that a high level of agreement would demonstrate the usefulness of the rubric descriptions in discriminating among levels of performance in the four categories of reading behavior included in the rubric.

The results of the comparison of scores suggest a high rate of agreement between expert and novice scorers. In 16 of 30 cases, the experts arrived at scores identical to those of the novices; in 12 additional cases, the expert score differed by only a single point. In the remaining 2 cases, the expert's score differed from the novice's score by two points. Thus it seems that the descriptions of differing point values associated with the four dimensions of fluency assessed on the rubric are very clear and enable users to discriminate among distinct levels of fluency. We concluded that the CRI Fluency Rubric can be used with confidence and consistency even by inexperienced users.

Validity

Our objective in creating a fluency rubric was to arrive at a numerical expression that could provide CRI users with an immediate and accurate appraisal of a reader's fluency. Validation of the Fluency Rubric was complicated, however, by our insistence that fluency be assessed only at the reader's instructional level. We believe that we can obtain only distorted notions of a reader's fluency if we base our assessment on materials that the reader finds too easy to read or too difficult. As a consequence, we were unable to simply correlate the fluency score with the reader's comprehension at that same grade level. The fact that the instructional level represents by definition a constricted range of comprehension scores would skew any attempt to calculate a correlation coefficient.

Instead, we reasoned that because fluency has been shown to be a fairly consistent correlate of reading achievement, a valid fluency score should be related to a reader's grade-level achievement in reading. Consequently, we compared the highest instructional level identified by the CRI with each reader's grade level. Our assumption was that, if the rubric is working as it should, readers with a higher fluency score on materials that present some challenges should be achieving at a higher level in reading than dysfluent readers. A child who was instructional at fourth grade but who was attending second grade would receive a score of +2; a child whose instruction level was third grade but who was attending fifth grade would receive a score of −2. We then correlated the level of fluency based on the CRI Fluency Rubric to the deviations of grade level and instructional level.

Instructional levels could be calculated for 208 cases and the overall bivariate correlation between CRI fluency scores and grade-level achievement in reading was .67 ($p < .0001$). This correlation was high enough to demonstrate that fluency scores are a valuable dimension of reading assessment and that they are clearly related to reading achievement. At the same time, the correlation is low enough to demonstrate that fluency cannot serve in isolation from other factors as a sufficient measure of reading proficiency.