

DEVELOPMENTAL READING ASSESSMENT

Note: This review discusses three assessment instruments: the Developmental Reading Assessment, Kindergarten through Grade 3, Second Edition (DRA2, K–3, Beavers, 2006), the DRA Word Analysis (Beavers, 2005), which is included in the DRA2, K–3 kit, and the Developmental Reading Assessment, 4–8, Second Edition (DRA2, 4–8; Beavers & Carter, 2003). Information for the technical adequacy section of this review was obtained from the DRA2 and Word Analysis teacher guides, the technical manual for the original DRA K–8 (Celebration Press/Pearson Learning Group, no date), two technical reports for the original DRA K–8 available on the publisher’s web site (Weber, 2000; Williams, 1999), and a one-page summary of DRA2 technical updates obtained from the publisher. According to the publisher, a technical manual for the DRA2 will be available by the winter of 2007, with analyses from data collected during field tests in the spring and fall of 2006. There are no technical materials for the DRA Word Analysis.

OVERVIEW

The Developmental Reading Assessment (DRA) is a set of individually administered criterion-referenced reading assessments for students in kindergarten through Grade 8. Modeled after an informal reading inventory, the DRA is intended to be administered, scored, and interpreted by classroom teachers. There are two instruments in the DRA series: the Developmental Reading Assessment, Kindergarten through Grade 3, Second Edition (DRA2, K–3, Beavers, 2006), which includes the DRA Word Analysis (Beavers, 2005), and the Developmental Reading Assessment, 4–8, Second Edition (DRA2, 4–8; Beavers & Carter, 2003). The DRA2 K–3 and 4–8 are intended to identify students’ independent reading level, defined as a text on which students meet specific criteria in terms of accuracy, fluency, and comprehension. Additional purposes include identifying students’ reading strengths and weaknesses, planning instruction, monitoring reading growth, and, for the DRA2 4–8, preparing students to meet classroom and testing expectations and providing information to stakeholders regarding reading achievement levels. The DRA Word Analysis is designed to evaluate the phonological awareness and phonics skills of students in kindergarten and early first grade and the word analysis skills of below-grade-level readers in Grades 1 through 5.

Materials for the DRA2, K–3 include a spiral-bound teacher guide, 23 leveled texts for kindergarten through Grade 1, three copies each of 22 leveled texts for Grades 2 through 4, a spiral-bound blackline masters book, a blackline masters CD, a laminated assessment procedures overview card, a training DVD, a timing clipboard with calculator, an organizer with hanging folders, and 30 student assessment folders, packed in a storage box. Additional Word Analysis materials include a spiral-bound teacher guide containing blackline masters, a spiral-bound stimulus book, and a training CD. Examiners must provide coins or counters for one phoneme segmentation task. Materials for the DRA2, 4–8 include a spiral-bound teacher guide, three copies each of 28 leveled texts, a spiral-bound blackline masters book, a blackline masters CD, a training DVD, a laminated assessment procedures overview card, a timing clipboard with calculator, an organizer with hanging folders, and 30 student assessment folders, packed in a storage box. The 4–8 assessment books consist of 20 texts for on-level

readers in Grades 4 through 8 and 8 “bridge pack” texts for intermediate and middle school students reading at a second- or third-grade level.

Changes to the DRA2, K–3 include timing oral reading beginning at level 14 (middle of Grade 1), having students respond in writing rather than by retelling beginning at level 28 (end of Grade 2), and the addition of a vocabulary item to the comprehension questions. Changes to the DRA2, 4–8 include combining teacher guides for 4–8 and bridge pack texts, revision or replacement of many bridge pack texts, and addition of a notetaking page for levels 60, 70, and 80 (Grades 6 through 8) for use in the written summary.

ASSESSMENT TASKS

The DRA2, K–3 and 4–8 each consist of two basic components: (a) a student reading survey and (b) a set of leveled paperback books, each with a reproducible text-specific teacher observation guide and student booklet. Except at level A, which has only one text, each level includes two fiction texts. Two additional nonfiction texts are included at levels 16, 28, 38, and 40 of the DRA2, K–3 and at each level of the DRA2, 4–8. The DRA Word Analysis consists of 40 tasks organized into five skill areas or “strands.” Table 1.1 lists the DRA tasks by skill area, beginning with the five essential reading components identified by the National Reading Panel (NRP, 2000) and endorsed in Reading First.

TABLE 1.1 DRA Tasks by Component/Skill Area

Component/ Skill Area	DRA Word Analysis Tasks	DRA2, K–3 Tasks	DRA2, 4–8 Tasks
Phonemic awareness	Rhyming (2 tasks) Alliteration (2 tasks) Segmentation (3 tasks) Phonemic awareness (5 tasks)		
Alphabetic principle/ Phonics	Letter naming (2 tasks) Word list reading (4 tasks) Spelling (5 tasks) Decoding (5 tasks) Substitutions/analogies (4 tasks) Structural analysis and syllabication (6 tasks)		
Fluency		Oral reading fluency (words per minute for contextual reading beginning at level 14)	Oral reading fluency (words per minute for contextual reading)
Vocabulary		Retelling/Scaffolded Summary: Vocabulary (one of seven	

		comprehension items)	
Comprehension		Oral or written responses to open-ended comprehension questions	Written responses to open-ended comprehension questions
Print concepts	Metalanguage (2 tasks)		
Reading engagement		Oral or written responses to a student reading survey	Oral or written responses to a student reading survey

ADMINISTRATION

The DRA2, K–3 and DRA 4 –8 can be administered on an annual or semiannual basis or more frequently in the case of struggling readers. The student completes a reading survey (orally for younger and less proficient students and in writing for levels 28 and up), after which the teacher conducts an individual reading conference, which includes a prediction component and oral reading of a designated section of the text or the entire text for younger students. While the student reads aloud, the teacher uses a text-specific observation guide to record nine categories of reading behavior, including six types of errors: (a) substitutions, (b) omissions, (c) insertions, (d) reversals, (e) incorrectly sounded out words, and (f) words told by teacher. If a student has five or more miscues, the teacher completes an oral reading analysis, which involves copying each substitution and noting the number of miscues and teacher-supplied words and the types of decoding strategies and miscues.

The total number of oral reading errors is converted to an accuracy score. Beginning at level 14, oral reading is timed, using a words-per-minute (WPM) metric. This reviewer recommends calculating reading fluency using the words-correct-per-minute (WCPM) metric, which has been validated as a predictor of reading proficiency, including performance on high-stakes assessments, across a wide variety of student populations (see Kame’enui & Simmons, 2001, for a review).

Three to four aspects of oral reading are rated on a 4-point scale, with the features rated varying by text level (e.g., expression, phrasing, rate, and accuracy for students reading at a Grade 4 to Grade 8 level). If the student’s score for rate or accuracy falls in the emerging/intervention (i.e., lowest) range for that text, the assessment is stopped, and a lower level text is administered at another date. Students demonstrate comprehension by retelling the story or text for levels 4 through 24 and by responding in writing to questions and prompts in a booklet for levels 28 and above. Students are not permitted to refer to the text during the prediction component, but they may use it to answer comprehension items.

The DRA Word Analysis is designed to be administered in the beginning of the year after the DRA has been administered or at midyear for kindergarten students. For students reading below grade level because of poor word analysis skills, the author recommends administering both the DRA and Word Analysis three times a year for progress monitoring and instructional planning purposes. The teacher guide includes a table with criteria for determining which students should take the assessment, based on beginning, mid-year, and

end-of-year independent DRA text level, and a second table lists a set of suggested tasks based on current DRA text level.

There are no time limits for the DRA2, K–3 and 4–8. Teacher guides estimate 10 to 15 minutes for the student reading survey, 6 to 20 minutes for the one-to-one conference, and 30 to 45 minutes for the silent reading and written components, depending on level. Estimates are based on students who are reading on grade-level, however, and struggling readers are likely to require more time. Each Word Analysis task takes about 2 to 3 minutes, with somewhat longer times for higher level spelling tasks. Several tasks have item and/or task time limits, including letter naming, sound naming, and single word reading measures. The teacher guide notes that most emergent readers in the field tests completed their task sets in less than 12 minutes but that some students required multiple administrations. Low-performing students are likely to require more time, especially for tasks requiring written responses.

Several aspects of administration for the DRA2, K–3 and 4–8 are problematic and compromise both reliability and validity. First, text selection is based on teacher judgment rather than on an objective, standardized routing task, such as a graded word list or set of screening measures. Instead, teachers are instructed to select texts on which they believe students can achieve an independent reading level. No theoretical rationale or empirical data are provided in support of this procedure, which differs from that of other large-scale reading assessment batteries (e.g., Texas Primary Reading Inventory [TPRI; Foorman et al., 2002-2003]; Phonological Awareness Literacy Screening [PALS; Invernizzi, Meier, & Juel, 2003-2005]), and reading inventories (e.g., Qualitative Reading Inventory, IV [Leslie & Caldwell, 2006]). Teacher guides recommend assessing with fiction in the fall and nonfiction in the spring, which introduces additional variance due to lack of control for genre. The training DVD for the DRA2, K –3 indicates that teachers should select texts that are on, at, and above the student’s reading level, although this is not explicitly stated in the teacher guide. Teacher guides include tables with suggested DRA text levels for readers on, at, and above grade level, as well as lists of comparable trade books relative to DRA levels, but the process remains highly subjective and vulnerable to the operation of confirmation bias (Nickerson, 1998). For example, teachers who believe that students are reading on a particular guided reading level may assess a student only with a DRA text corresponding to that level rather than selecting from a broader range of text difficulty (Invernizzi, Landrum, Howell, & Warley, 2005).

Second, further inconsistency in the text selection process arises from a student-choice component. After the teacher has selected three or four texts, students are then invited to choose a text that seems “just right.” Although there is some evidence that text choice can enhance performance on reading comprehension tasks (e.g., Reynolds & Symons, 2001), no empirical data evaluating differences in text level or reading performance for teacher versus student text choice are presented.

Third, teacher guides indicate that the DRA can be administered over several days. Although this guideline may be an effort to reflect the realities of today’s classrooms and the time-consuming nature of the DRA, comprehension scores may differ for students who read

the entire text and answer comprehension questions on the same day and those who complete the oral reading portion on one day and finish reading the selection and respond to questions on another day. It is noteworthy that in the Vermont adaptation of the DRA (VT-DRA, Vermont Department of Education, 2005), teachers are required to complete an assessment with one DRA text in a single testing session.

A fourth concern regarding DRA administration procedures relates to the vague guidelines for word supply during the oral reading component. Although the record of oral reading guidelines included in the teacher guides indicates that a “word told by teacher” is an error, no information is provided as to when the teacher is to supply a word to a struggling student (e.g., after a 3-second pause, after a 5-second pause, after the student has made an attempt to decode the word, etc.). Unfortunately, the training DVDs offer little help in clarifying this issue. On the DRA2, K–3 training DVD, the students sometimes pause for a considerable period while struggling with an unfamiliar word, but teacher response varies, with one teacher inviting the student to “try” and another remaining silent until the student continues. On the DRA2, 4–8 DVD, none of the teachers provide any words to the students, who make very few oral reading errors and make no requests for teacher assistance. Because differences in word supply procedures can have a significant impact on both reading rate and comprehension, guidelines should be standardized.

Some of the administration procedures for the DRA Word Analysis are also likely to increase construct-irrelevant variance in the results obtained. The teacher is instructed to stop the assessment when the student demonstrates “some control,” “little control,” or “no control” on any three tasks or “is becoming distracted.” Distraction is not operationally defined, and teacher definitions of and thresholds for tolerating distraction are likely to vary. For subsequent administrations, only those tasks on which the student did not demonstrate control are readministered. Because “control” is defined as 100% success on a task, this implies that teachers should continue to readminister a task until the student achieves 100% accuracy, even for tasks with a large number of items. No rationale or empirical data are provided in support of the score ranges or the discontinuation rule. Moreover, because each task has only one form, improvement may reflect the operation of practice effects as well as genuine growth in phonological awareness and word analysis skills.

SCORES

The authors assert that teachers can analyze student performance in 5 to 10 minutes, an estimate appears highly optimistic for an assessment requiring so many subjective judgments. If teachers audiotape oral reading and retellings for later review, as suggested, this will greatly increase scoring time. Teachers will also need to invest a substantial amount of time and effort to master the miscue coding system. Teacher observation guides present directions, questions, and prompts for each book, including story or text overviews for use in scoring retellings on the DRA2, K–3. Responses are scored using a rubric, termed a continuum, with descriptors of student behaviors in reading engagement, oral reading fluency, and comprehension. Responses on the survey and to comprehension items are scored only in terms of content, and structural and mechanical errors are ignored. There are two versions of the continuum – one for fiction and one for nonfiction. Performance on each item is rated on

a 4-point scale, corresponding to four categories or stages of reading performance: emerging/intervention, developing, independent, and advanced for the DRA2, K–3, and intervention, instructional, independent, and advanced for the DRA2, 4–8. Ratings are summed to determine overall performance in each of the three areas.

Despite the authors' assertion that the continuums include "consistent, clear criteria" for scoring student responses, many aspects of scoring make the DRA highly vulnerable to interrater variance. To evaluate responses on the student reading survey, teachers must be familiar with the books reported or listed by their students not only in terms of title and genre but also grade level with a high degree of precision (i.e., "2–3 titles slightly below grade level" = instructional level; "generally on-grade-level texts" = independent level; "many on- and above- grade-level texts" = advanced level). Answers are not provided for comprehension items, all of which are open-ended. Instead, teachers are directed to score items based on their knowledge of the text. Given the range of reading grade levels in typical upper elementary and middle school classrooms, this means that many teachers using the DRA2, 4–8 will need to read all 32 books, some of which are over 1000 words in length. Moreover, teachers must be able to remember the content of the books well enough to score six comprehension items on the 4-point rating scale, with small gradations between the four performance levels for many items. For example, for the Retelling: Vocabulary item on the DRA2, K–3, the descriptor corresponding to a rating of "developing" is "uses some language/vocabulary from the text; some understanding of key words/concepts," whereas the descriptor corresponding to a rating of "independent" is "uses language/vocabulary from the text; basic understanding of key words/concepts."

The ambiguity of the scoring guidelines is compounded by the relative lack of scoring examples for lower performing readers. The DRA2 K–3 teacher guide includes examples of independent and advanced responses to comprehension items for each text for levels 28 through 40 but no examples of comprehension responses at emerging/intervention or developing/instructional levels. The author justifies this omission on the grounds that teachers will be able to recognize partially correct responses if they understand what is included in independent and advanced responses. The DRA2, 4–8 teacher guide includes examples from a range of performance levels for the student reading survey as well as for comprehension items for three texts (a bridge pack level 34 fiction text, a level 50 nonfiction text, and a level 70 fiction text). Examples of independent responses to comprehension items are also provided for each of the 4–8 texts, as well as examples of both independent and advanced responses for all eight bridge pack texts.

For the DRA Word Analysis, scores are dichotomous for all items. Responses to spelling tasks are scored for the presence of developmentally appropriate features rather than whole-word correctness, which increases task sensitivity. Subjective judgment is required to evaluate many of the items. For example, on Task 28, the student is required to provide an oral sentence to indicate understanding of a word with a suffix. If the teacher judges that the student's sentence does not indicate understanding, the student is asked to define the word. Examples of acceptable and unacceptable responses are not provided for the items, many of which do not lend themselves readily to definitions (e.g., *himself*, *couldn't*). Moreover, although the record sheet includes space to write a word or two if the student's answer is

incorrect, no space is provided for recording sentences or definitions for later review. Scoring guidelines for letter sound tasks should be clarified. On the training DVD, one teacher comments during the discussion with the author that a student added a vowel in pronouncing a consonant sound (“suh” for /s/), implying that this is an error, and the author concurs, whereas the teacher guide does not indicate that the addition of a vowel to a consonant sound renders it incorrect.

INTERPRETATION

All of the DRA assessments are criterion-referenced, and no normative data are provided for interpretive purposes. Descriptor ratings for each item are summed to determine overall performance in each of the three areas assessed. Performance levels can be recorded on a graph for each DRA text level, with the goal of scoring within the independent range or higher in both fluency and comprehension on a grade-level text by the end of the school year. The DRA2, K–3 teacher guide includes a table with fall and spring text benchmarks for each grade (e.g., spring of Grade 1 = levels 16 to 18), but no data are provided in support of these criteria. Compared with the reading accuracy benchmarks commonly used to assign functional reading levels, DRA accuracy benchmarks involve four levels and are set higher. For example, for level 60 (Grade 6 readers), intervention-level accuracy is defined as 95% or less, instructional as 96%, independent as 97% to 98%, and advanced as 99% to 100%. As noted above, reading rate is based on a WPM rather than the preferred WPCM metric. Surprisingly, neither guide includes a complete case example to illustrate the scoring and interpretative process for a specific student or examples to demonstrate how the results can be used to identify reading strengths and weaknesses or plan instruction.

Scores for each Word Analysis task are converted to four levels of competency, termed “levels of control:” no/little control (0–39% correct), some control (40–79% correct), gaining control (80–99% correct), or control (100% correct). The teacher guide includes a form for recording correct and incorrect spelling examples and oral reading miscues (the latter from the DRA assessment) for analysis, but no guidelines or case examples are included to demonstrate how teachers can use the information for diagnostic or instructional planning purposes.

TECHNICAL ADEQUACY

Development and Field-Testing

The DRA was developed in 1986 in the Upper Arlington City School District in Ohio by a committee of teachers and educators, headed by Joetta Beaver, an early education teacher. The original 20-text DRA K–3 was field tested in spring of 1996 by 84 teachers with 346 students in kindergarten through Grade 3 in 10 states and a province in Canada. Sample-wide percentages are reported for race/ethnicity and gender, and the percentages of schools in each of three types of community is given, but the information is not grade-specific, and relevant U.S. census data are not provided for comparative purposes, making it difficult to evaluate sample representativeness even in terms of race and gender. A review of 2000 U.S. census data reveals that African-American, Hispanic, and Asian students were

underrepresented, although the degree of underrepresentation is difficult to specify because a sizeable proportion of students were not coded for race/ethnicity. No information is provided regarding other key variables, such as parent educational level, socioeconomic status, or disability status. Grade-level sample sizes were very uneven (*ns* of 10 to 162), there is no indication as to how many students read each of the 20 texts, and total sample size falls well below the criterion level (i.e., 1000 examinees) (Rathvon, 2004). A set of alternative texts were developed and field tested in May and June of 2000 in 18 states and two Canadian provinces in a sample of 208 students in kindergarten through Grade 3, with 157 teachers administering the original texts and the 20 alternative texts to each student. Grade-level sample sizes were again highly variable (*ns* of 19 to 101), there is no indication of how many students read each text, and sample characteristics are not reported by grade. A field test with a small sample described only in terms of number ($n = 95$) was conducted with the revised texts and forms in September of 2000. Additional revisions were made based on teacher feedback, but few details are provided.

The original DRA 4–8 texts were field tested in October and November of 2001 with 706 students in Grades 3 through 9 in 26 states and one Canadian province. The 181 teachers were asked to assess students using three different texts. Sample-wide rather than grade-specific percentages are reported for race/ethnicity, gender, and school community as before. Grade-level sizes were highly variable (*ns* of 4 to 271), indicating that only a few students read each text at some levels. A second field test was conducted in February and March of 2002 with 129 teachers and 761 students in Grades 4 through 8 in 20 states and Canada. Grade-level sizes were again highly variable (*ns* of 6 to 294). In the first field test sample, African-American, Hispanic, and Asian students were slightly underrepresented, whereas in the second sample, Caucasian students were overrepresented. It should be noted that for both field tests, teachers were instructed to include only students who were able to read their present grade-level text with 97% accuracy or better. This suggests that students with disabilities, who typically perform below grade level in reading, were not included in the sample. This is unfortunate because the exclusion of lower performing readers will result in artificially inflated mean scores and reduced variance.

According to a publisher representative, teachers participating in the DRA2 field test sample were instructed to select one student from each of the following categories: below grade level, slightly below grade level, on grade level, and above grade level. Despite the effort to sample a broader range of reading competencies in this edition, there is some indication that more proficient readers may still be overrepresented. In the DRA2 4–8 teacher guide, the authors indicate that there were no intervention-level responses in the field test for half of the comprehension items on two of the three texts used in the scoring examples (prediction, literal comprehension, and reflection for *Storm Chasers* [level 50] and prediction, literal comprehension/note taking, and interpretation for *Lost!* [level 70]).

Very little specific information is available about the development of the DRA Word Analysis. In spring of 2000, an analysis was conducted of all words in the DRA texts, the types of miscues made by students while reading DRA texts, and “research-based information” on developing readers’ word analysis skills. After an initial draft was shared with a group of classroom, reading, and speech teachers, the initial instrument was pilot tested

in central Ohio in spring and fall of 2002 in four schools. Based on teacher feedback, revisions were made in terms of directions, tasks, and assessment materials. Additional field tests were conducted in the winter and fall of 2003 by classroom and reading teachers in four types of school communities. Although the author asserts that field tests included “representative samples” in terms of race/ethnicity, gender, and grade-level, no specific information about sample characteristics is provided, even in terms of total number of students, and no specific information is provided about the nature of the revisions.

Reliability Evidence

Evidence of interrater reliability for the DRA K–3 is based on a study (Williams, 1999) with 306 students in kindergarten through Grade 3 (*ns* of 33 to 125) reading on text levels from A to 44 (preprimer to Grade 5). Eighty-seven teachers in 10 states conducted and audiotaped DRA conferences with three or more students, after which each tape was rated by a second and then a third teacher (total teachers = 127). Interrater agreement based on Rasch analyses for five rating scale items (accuracy, comprehension, reading stage, phrasing, and reading rate) was .80 for the first two raters, which is acceptable for measures designed for screening purposes but below criterion levels for instruments intended for individual diagnostic and programming purposes. Moreover, when all three raters were considered, interrater agreement fell to .74. It is not clear whether the second and third raters had access to the relevant texts when they reviewed the audiotapes, and information regarding the number of students at each grade and text level is not provided.

Two studies were conducted comparing teacher ratings with expert ratings, an approach that provides evidence of scorer accuracy as well as scorer consistency. Weber (2000) compared teacher ratings of oral reading accuracy with ratings of an expert who conducted DRA conferences with four students who completed between two and four DRA text levels (levels A through 44) while 10 teachers observed behind a one-way mirror. Percent agreement with the expert was 100% for a 3% to 5% agreement category and ranged from 70% to 100% for a 2% agreement category. Because differences between DRA accuracy categories are very small, however, even a 2% disagreement can represent a difference in performance levels. Moreover, teacher-expert agreement for the many fluency and comprehension scores that require a greater degree of subjective judgment was not evaluated.

In a second study comparing both teacher-teacher and teacher-expert ratings (Fisher, 2003), five students were videotaped during a DRA conference (levels A, 1, 12, 40, and 70) with 44 teachers rating the three elementary grade students but not the two middle school students. Results of Rasch analyses indicated that facets for items (summed score for individual teacher ratings) and students were highly reliable (*rs* = .92 and .99, respectively), whereas the facet for rater displayed very low reliability (*r* = .27). It is not clear why the middle school students were not rated, and the findings are not likely to increase users’ confidence in the consistency of the DRA across examiners. In a second phase of the study, teachers (*ns* = 7 to 30 per student) completed the DRA observation guide for six students (one each in kindergarten and Grades 1, 2, 3, 4, and 7), who completed between one and three DRA levels, with five or six subscores evaluated per student. Percent agreement with an

expert was highly variable, with generally high levels of agreement for phrasing and fluency but much lower levels of agreement for other scores, especially comprehension and miscues. Perfect agreement for overall reading stage or comprehension level was also (57.1% to 70%).

Test-retest reliability coefficients (approximately 3-week interval) for independent reading level in a sample (Weber, 2000) of 306 students in Grade 1 through Grade 3 ($n_s = 100$ to 104) were high for all three grades ($r_s = .92$ to $.99$), but it is unclear whether students were tested twice on the same text, on alternative texts within the same level, or on texts from different levels. Moreover, the extent of practice effects cannot be adequately evaluated because mean scores are not reported by text level. Standard deviations were also large, especially for Grade 1 students, indicating lack of measurement precision. No stability estimates are reported for the DRA 4–8. Stability estimates should be reported by text level for each component of the DRA, including accuracy, rate, and comprehension scores, as well as for overall reading level.

Internal consistency data collected during the Williams (1999) study cited above indicated high levels of consistency for the five items across all three raters (Cronbach's $\alpha = .98$) and for DRA texts ($.97$), but no other details are provided. Additional evidence of internal consistency is needed, especially in view of the fact that only a single text may be administered. In addition to alpha coefficients, means, standard deviations, and standard errors of measurement should be reported for subscores and overall score because of the restriction of range arising from students reading a limited number of text levels. Internal consistency estimates should also be provided for gender, racial/ethnic, and disability subgroups to demonstrate that the DRA is equally reliable for all examinees and contains little or no bias relative to these groups.

Because the DRA is designed to determine students' reading level rather than to rank students relative to their age or grade peers, evidence of alternate-form reliability (i.e., equivalence of texts within a level) is critical in establishing the reliability and validity of the results. Although the technical manual indicates that several alternate texts were revised based on field test data and teacher feedback and to match the word-count criterion, the only specific information regarding alternate-form equivalence is number of words per text, which is listed in the reading rate charts at the back of the teacher guides, and the readability data for the original and alternate K–3 texts obtained in the Lexile study described below in the content validity section. Mean scores, standard deviations, and standard errors of measurement should be presented to demonstrate that alternative texts at each level are of similar difficulty. In addition, the percentage of students obtaining the same overall scores in oral reading fluency and comprehension and the identical performance level should be presented for the alternate texts at each level.

No reliability evidence of any kind is presented for the DRA Word Analysis. Given the vulnerability of phonological awareness measures to both examiner and scorer variance (Rathvon, 2004), studies of internal consistency, stability, and interrater reliability should be a priority.

Validity Evidence

Content validity evidence

According to the author, the DRA was designed to reflect the characteristics of good readers as observed by teachers and reported in the research literature. The running record format was modeled on Clay's *Observational Survey* (Clay, 1993). The theoretical rationale includes a review of the premises underlying the DRA, with citations from the literature. The technical manual also reports the results of teacher surveys (*ns* of 80 to 175) conducted after the field tests in which teachers responded to a variety of statements about the assessment materials, their utility, and other dimensions. Although teachers agreed that the DRA was helpful in describing reading behavior and identifying instructional goals, the results reveal consistent concerns regarding the adequacy and accuracy of the comprehension assessment and the accuracy of text leveling. Moreover, return rates were as low as 46%, reducing the generalizability of the findings.

Although the technical manual and teacher guides emphasize the involvement of committees of teachers in the development and revision of the DRA, there is no evidence that external reviewers, such as curriculum, reading, and assessment experts, participated in the development, revision, or validation process. Nor is there any indication that the authors reviewed any assessments other than Clay's Observation Guide in developing the original or revised DRA K–8. Similarly, there is no indication in the Word Analysis teacher guide that any of the numerous phonological awareness and decoding measures available from commercial publishers or in the research literature were reviewed during development, and none are cited in the reference list, which consists of primarily of instructionally oriented texts. In the study by Fisher (2003) cited above, significant point-biserial correlations were found between 10 DRA individual items, including prediction, phrasing, and accuracy, and total score, but only a single coefficient is reported ($r = .40$). No studies of item discrimination, item difficulty, or differential item functioning are reported. Moreover, no data from field test samples comparing the effects of various assessment and response formats on performance (e.g., answering comprehension questions with and without consulting the text, responding orally or in writing to comprehension questions, silent versus oral reading, illustrated vs. print-only texts, timed vs. untimed Word Analysis tasks, etc.) are presented.

Evidence of lack of bias is also very limited. The authors state that the texts were developed to reflect cultural diversity and to include “strong” female and male characters, but there is no evidence that judgmental procedures, such as task and item reviews by curriculum experts and representatives of various demographic group, or statistical procedures, such as Item Response Theory (IRT) analyses, were conducted to evaluate possible task or item bias. Differential item functioning (DIF) studies are especially important in view of a study (Buchanan, 2002) reported in the technical manual indicating that Caucasian students demonstrated significantly greater increases in DRA independent level compared with African-American and Hispanic students.

Because student performance is keyed to text level, the appropriateness of text selection and the accuracy of the leveling are critical validity issues. Texts were taken from a variety of sources or were written specifically for the assessment. The technical manual and

teacher guides list the criteria used to determine difficulty level for fiction and nonfiction texts, but specific data for individual texts are not presented. The Fry readability index was used to level texts at level 30 and above, with analyses extended to all 100-word samples in each text. Texts were included as long as they fell within one grade level of the DRA level for which the text was designed. Two texts were relevelled and modified based on Dale-Chall readability measures. In September 2003, the Lexile framework was used to evaluate the readability of each DRA text. Results are presented in terms of genre, mean sentence length, mean word frequency, and Lexile measure for the DRA K-3 original assessment texts, K-3 alternative assessment texts, and DRA 4-8 assessment texts, but the information is in three separate tables, making it difficult to conduct within-level comparisons for the K-3 texts. Moreover, results for the Fry and Dale-Chall analyses discussed above are not presented. Because different formulae can yield very different estimates for the same text, multiple readability formulae should be applied and presented in a single table, and the decision rule for assigning text level explicitly stated. No theoretical rationale or empirical data are offered in support of the omission of a standardized task to estimate student reading level.

The technical manual cites data from the Weber (2000) study discussed above to demonstrate that accuracy rates decrease as text level increases, but the sample is too small ($n = 4$) for generalization purposes, and one of the four students did not demonstrate the expected pattern. Means, standard deviations, and standard errors of measurement should be presented for accuracy, rate, and comprehension scores for field test students reading adjacent text levels to document level-to-level progression.

According to the technical manual, DRA 4-8 reading rate categories were based on data from 50 to 60 students collected by the authors. In the DRA2 series, WPM ratings are based on the National Assessment of Educational Progress (NAEP) 1995 oral reading fluency study (U.S. Department of Education, 1995) as well as field test data, but no specific data are provided in support of the WPM ranges. Accuracy rates for the DRA2 are somewhat higher than for the original DRA. According to a publisher representative, accuracy percentages were adjusted based on field test data to ensure that students were able to comprehend the texts adequately, which translated into higher accuracy rates, but no specific information is provided relative to the cut scores for the various performance levels.

Very little information is available regarding the development, piloting, or validation of the DRA Word Analysis. A table in the teacher guide shows how tasks in each strand are distributed across subskill areas and reflect the behavior of proficient readers, but no theoretical or empirical basis is presented for test organization or for the content and format of the tasks, many of which use atypical item types. Because variations in format, content, and item type can have a significant impact on estimates of students' phonological awareness and decoding skills (Rathvon, 2004), the author should provide an explicit rationale for the selection of task formats and item types, accompanied by citations from the research literature.

Criterion-related validity evidence

Evidence of validity based on the DRA's relationship to other measures of reading skills is available only for primary grade students. In a study (Weber, 2000) with 284 students in Grades 1 through 3 in four elementary schools, correlations between DRA K-3 independent reading level and grade equivalents for the Comprehension subtest on the Iowa Tests of Basic Skills were generally in the moderate range (.54 to .84). For a sample of Grade 2 students ($n = 2470$) from a large urban/suburban district in Fort Bend, Texas (Williams, 1999), DRA independent level assessed at the end of the 1998-1999 school year was moderately correlated with fall of Grade 3 normal curve equivalent (NCE) scores on the Iowa Test of Basic Skills Vocabulary and Reading Comprehension subtests and for Total Reading ($r_s = .68, .68, \text{ and } .71$, respectively). No concurrent validity evidence is presented documenting the relationship between the DRA and standardized or criterion-referenced tests of reading, vocabulary, language, or other relevant domains for students in kindergarten or Grades 4 through 8. Studies examining the extent to which individual students obtain identical performance levels on the DRA and validated reading measures are especially needed. No information is provided to document the relationship between the DRA Word Analysis and any criterion measure. No concurrent validity evidence is presented for any of the DRA assessments in terms of the relationship between DRA performance and contextually relevant performance measures, such as teacher ratings of student achievement or classroom grades.

Construct validity evidence

As evidence of the developmental progression of the texts, the technical manual includes a brief summary of informal studies conducted by the author for students reading three adjacent text levels, but no details are provided. In a sample of Grade 2 students (number unspecified) in one elementary school in Vermont from 1998 through 2000 using a modified version of the DRA (VT-DRA), there was a positive relationship between performance/achievement levels and DRA levels over the 3-year period. Results from Louisiana statewide DRA administrations for spring of 2000 through 2002 for students in Grades 1 through 3 ($n_s = 4,162 \text{ to } 74,761$) show an increase in DRA levels across grades, as well as changes in DRA level for a matched sample of student ($n = 32,739$) over the three-year period, indicating that the skills being measured are developmental in nature. As evidence of the DRA's ability to detect changes in reading performance due to intervention, the technical manual reports changes in reading level and reading stage by mean score and percent per category for students participating in a summer school remedial reading program in six Austin, Texas school districts in 1999 and 2002 ($n_s = 1,101 \text{ and } 1,994$, respectively). Mean gains in text reading level were 2.0 and 1.8, respectively.

As evidence of discriminant validity, the technical manual reports data from a 2000-2001 study in Louisiana (Buchanan, 2002) that compared changes in DRA independent reading level by race for a large sample of students ($n = 93,668$) from fall 2000 to spring 2001. Although Caucasian students demonstrated greater increases in reading level compared with African-American and Hispanic students, effect sizes accounted for less than 1% of the variance of the change.

Two studies evaluating the relationship between Lexile Scale measures and DRA independent reading level are reported as evidence that the DRA running-record format is a valid method of assessing reading comprehension. In a 1998 study with a sample of Grade 2 and 3 students ($n = 259$) with DRA independent reading levels from level 9 to level 30, the overall correlation between Lexile measure and DRA level was .69. Correlations are not reported by level or grade, however, and the correlation may be spuriously high because results are combined across levels. In a 2003 study with 562 students in Grades 1 through 3, correlations between Lexile reading test means and DRA means ranged from .73 to .81 across the three grades. Inspection of the graph depicting the relationship between DRA level and mean Lexile measure reveals that Lexile measures increase as DRA levels increase but also a high degree of variability within and between DRA levels. For example, Lexile measures corresponding to DRA level 25 range from approximately -200 to approximately 900. For a subsample ($ns = 138$ to 222, number per grade not specified), correlations between Lexile measure and DRA subscores were moderate for accuracy (.53), but low for comprehension (.26). Moreover, the correlation between DRA accuracy and DRA comprehension was also unexpected low (.19). Task intercorrelations should be reported by grade and preferably by level for accuracy, rate, and comprehension.

Because the DRA Word Analysis is designed to identify children in need of additional interventions, evidence of diagnostic utility, such as demonstrated by group and intervention differentiation studies, should be presented, as well as evidence of its ability to identify individual children with reading problems (i.e., sensitivity and specificity indices).

USABILITY

Teacher and student materials are very attractive. The texts, which include full-color illustrations, should be very appealing to students, and efforts have been made to depict individuals in various gender and racial groups and in nonstereotypical ways. Learning to administer, score, and interpret the DRA requires a major investment in terms of time and effort, however, and the brief tables of contents in the teacher guides make it difficult to locate information quickly. The DRA2 K-3 and 4-8 training DVDs depict reading conferences with students at several DRA stages, including post-conference discussions between the teacher and one of the authors, but offer surprisingly little specific information in terms of scoring or interpretation. Moreover, a considerable amount of time on each DVD is devoted to a conversation between the two authors and an administrator about districtwide use of the DRA, whereas information relative to fine points of administration and scoring, including a demonstration of the miscue coding system, would have been more useful. The DRA Word Analysis training CD is very well done and includes task demonstrations with students at three different reading stages. No information is provided in the guides to assist teachers in determining what, if any, adaptations in administration, scoring, and interpretation are appropriate for assessing students from diverse linguistic or cultural backgrounds, other than a note in the oral reading guidelines on scoring variations in dialect.

The one-page directions for the timing clipboard need to be expanded to include more information on resetting the timer. The optional data management system available for both series permits direct score entry and generates a variety of reports that can be used for

instructional grouping, reporting individual student progress, and presenting classroom, school, and district results. Also available is a pen with a digital camera that converts text written on a special teacher observation guide into digital media so that data can be directly uploaded into a computer. A Spanish version of the second edition of the DRA that covers kindergarten through Grade 6 is also available. Users have access to a companion web site (www.connect2DRA.com) that includes information on updates and professional development.

LINKS TO INSTRUCTION

The DRA is designed to inform instruction within a guided reading model. Independent reading levels yielded by the DRA are correlated with the Fountas/Pinnell level system (Fountas & Pinnell, 1996). Teacher guides include a checklist of instructional activities based on the same categories as those in the DRA continuum and a class profile form for use in grouping students for instruction. The blackline master books include a section entitled “Moving into Instruction” with suggestions for instructional activities based on reading competencies at various DRA levels and a set of generic blackline masters related to DRA comprehension items, such as prediction, notetaking, and summarizing. The DRA Word Analysis teacher guide includes a useful set of mini-lessons and a list of recommended teacher references for each strand.

RELEVANT RESEARCH

The DRA has been used in several studies to evaluate the effects of reading intervention programs (e.g., Donis-Keller, Saunders, Wang, & Weinstein, 2004; Haenn, 2002), including one study cited in the technical manual (Curry & Zyskowski, 2000). In a concurrent validity study with 197 students in Grades 1 through 3 reported in the Phonological Awareness Literacy Screening–Grades 1-3 (Form B) Technical Reference (Invernizzi, Meier, & Juel, 2003), DRA instructional level was highly correlated with the spring 2001 PALS summed score (a combination of word list reading and spelling) ($r = .82$, $p < .01$). For a subsample of 96 students, DRA independent level and PALS summed score were also strongly related ($r = .81$, $p = .01$). The 2006–2007 school year is the last year in which DRA passages may be used as alternative passages to the PALS oral reading passages in the Virginia statewide screening program, however. Information regarding this decision will be posted on the PALS web site (www.pals.virginia.edu) in the near future (personal communication, July 28, 2006, Jennifer Howell, Ph.D., Research Scientist, K–3).

TEST REVIEWS

No reviews of any of the editions of the DRA could be located.

SOURCE AND COST

The DRA series is available from Pearson Learning (www.pearsonlearning.com). The price of the DRA2 K-3 comprehensive package is \$305.50. Cost for the DRA2 4-8

comprehensive package is \$245.50. Cost for the optional data management system varies from \$79.95 to \$384.95, depending on the package selected.

SUMMARY

The Developmental Reading Assessment, Second Edition (DRA2), is an attractive reading battery modeled after an informal reading inventory, with authentic texts, instructionally relevant measures of fluency and comprehension, and results that are meaningful to classroom teachers, parents, and other stakeholders. Although it is likely to be very appealing to educators, it also illustrates many of the concerns that have been expressed regarding the reliability and validity of informal reading inventories (e.g., Invernizzi et al., 2005; Spector, 2005). In terms of the original DRA, documentation of the adequacy of content coverage is limited, stability estimates are available only for Grades 1 through 3, internal consistency estimates are available only for kindergarten through Grade 3, and there is very little evidence of text equivalence within levels. Evidence of interrater reliability is unimpressive, even for overall reading level, and is available for only about half of the grades covered by the test. Although there is some encouraging evidence of the DRA's utility in predicting future reading achievement and responsiveness to intervention for primary grade students, very little evidence of criterion-related validity is available for older students. Moreover, although efforts have been made to clarify administration and scoring procedures in this edition, the text selection process and many aspects of scoring on the DRA2 remain highly vulnerable to inconsistency.

At this point, the DRA places high demands on teacher judgment in administration and scoring without also providing sufficient evidence that teachers can select texts aligned with students' actual reading level or achieve acceptable levels of scorer consistency and accuracy. This is a critical issue because without reliable scoring across examiners, establishing other types of reliability and validity is problematic, and users cannot be certain that changes observed in student performance are the result of genuine gains in reading proficiency rather than the result of measurement error.

Another set of concerns relates to the representativeness and relevance of the field test samples. Field test samples for the original DRA are incompletely described, have insufficient numbers of students at several grade levels and unknown numbers at each text level, and included only students reading on grade-level. Although sample characteristics for the new edition have not yet been made available, there is some evidence to suggest that lower performing students may be underrepresented, and there is no indication that students with identified disabilities were included. Moreover, there is currently no evidence that students with similar levels of reading ability perform similarly on the DRA, regardless of their membership in various demographic groups.

The apparent lack of participation by external reviewers in the development, revision, and validation of any of the DRA series is also troublesome. Although teacher input and feedback play an important role in validating classroom-based assessments, especially in terms of enhancing usability and ecological validity, outside review in the form of external evaluators and advisory review panels is essential to ensure content quality and

representativeness, assess psychometric soundness, and minimize the potential for bias or stereotyping.

According to the summary of technical updates obtained from the publisher, additional studies of text comparability within levels, text scaling across levels, alternate-form reliability, and internal consistency, including IRT analyses to equate texts and evaluate item bias, are being conducted, and the updated technical manual will be available in the fall of 2006. Although data from these analyses will be very helpful in evaluating the DRA2's psychometric characteristics, potential and current users must always consider the adequacy of the evidence in light of the uses they will make of the results, including the consequences of those uses. As stated in the latest edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999): "Precision and consistency of measurement are always desirable. However, the need for precision increases as the consequences of decisions and interpretations grow in importance" (pp. 29-30).

Until more information on the DRA2's psychometric characteristics becomes available and can be evaluated, DRA users cannot be confident that students will obtain the same reading levels with different teachers, on different occasions, and on different texts within the same level, or that the obtained performance levels are accurate reflections of students' actual levels of reading proficiency. Consequently, the use of the DRA for summative assessment or individual programming purposes cannot be recommended until additional reliability and validity evidence is provided. No reliability evidence of any kind is available for the DRA Word Analysis. Moreover, no evidence of its criterion-related or construct validity is currently available, such as information concerning its relationship to other validated reading measures, teacher ratings, or school grades; utility in predicting future reading achievement or response to intervention; or ability to distinguish between poor or proficient readers. In the absence of this evidence, it cannot be recommended for screening, diagnosis, progress monitoring, or intervention planning purposes.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.
- Beaver, J. M. (2006). *Teacher guide: Developmental Reading Assessment, Grades K– 3, Second Edition*. Parsippany, NJ: Pearson Education, Inc.
- Beaver, J. M. (2005). *Teacher guide: Developmental Reading Assessment Word Analysis*. Parsippany, NJ: Pearson Education, Inc.
- Beaver, J. M., & Carter, M. A. (2003). *Teacher guide: Developmental Reading Assessment, Grades 4–8, Second Edition*. Parsippany, NJ: Pearson Education, Inc.
- Buchanan, T. K. (2002, April). *Developmental Reading Assessment: Student achievement*. Study conducted for the Louisiana Department of Education, Division of School Standards, Accountability, and Assistance.

- Celebration Press/Pearson Learning Group. (no date). *DRA K–8 technical manual: Developmental Reading Assessment (DRA). Evaluaci'on del desarrollo de la lectura (EDL)*.
- Clay, M. M. (1993). *An observation survey of early literacy achievement*. Portsmouth, NH: Heinemann.
- Curry, J., & Zyskowski, G. (October, 2000). *Summer opportunity to accelerate reading (S.O.A.R evaluation)*. Office of Program Evaluation, Austin Independent School District, TX. ERIC document ED450141.
- Donis-Keller, C., Saunders, T., Wang, L., & Weinstein, M. (January, 2004). *Second year evaluation report for the Cornerstone Literacy Initiative*. Institute for Education and Social Policy, New York University. ERIC document ED486209.
- Fisher, W. A. (2003, December). *Reliability report*. Unpublished manuscript.
- Foorman, B. R., Fletcher, J. M., Francis, D. J., Carlson, C. D., Chen, D.-T., Mouzaki, A., et al. (2002). *Texas Primary Reading Inventory–2002-2003*. Center for Academic and Reading Skills. University of Texas-Houston Health Science Center & University of Houston.
- Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading*. Portsmouth, NH: Heinemann.
- Haenn, J. F. (April, 2002). Class size and student success: Comparing the results of five elementary schools using small class sizes. ERIC document ED486209.
- Invernizzi, M., Landrum, T. J., Howell, J. L., & Warley, H. P. (2005). Toward the peaceful coexistence of test developers, policymakers, and teachers in an era of accountability. *The Reading Teacher*, 58, 610-618.
- Invernizzi, M., Meier, J., and Juel, C. (2003). *Phonological Awareness Literacy Screening–Grades 1-3 (Form B) technical reference*. Charlottesville, VA: Curry School of Education, University of Virginia Press.
- Kame'enui, E. J., & Simmons, D. C. (Eds.). (2001). The role of fluency in reading competence, assessment, and instruction: Fluency at the intersection of accuracy and speed [Special issue]. *Scientific Studies of Reading*, 5(1).
- Leslie, L., & Caldwell, J. (2006). *Qualitative Reading Inventory, Fourth Edition*. Parsippany, NJ: Pearson Education.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 175-220.
- Rathvon, N. (2004). *Early reading assessment: A handbook for practitioners*. New York: Guilford Press.
- Reynolds, P. L., & Symons, S. (2001). Motivational variables and children's text search. *Journal of Educational Psychology*, 93, 14-23.
- Spector, J. (2005). How reliable are informal reading inventories? *Psychology in the Schools*, 42, 593-603.
- U. S. Department of Education. (1995). National Center for Educational Statistics. *Listening to Children Reading Aloud: Oral Fluency*, 1(1). Washington, DC.
- Vermont Department of Education (October, 2005). *Vermont Developmental Reading Assessment (VT-DRA) administration manual*. Retrieved July 23, 2006 from http://www.state.vt.us/educ/new/html/pgm_curriculum/literacy/dra.html.

- Weber, W. A. (2000). Developmental Reading Assessment and Evaluaci'on del desarrollo de la lectura: A validation study. Retrieved July 23, 2006 from http://pearsonlearning.com/correlation/rsp/ResearchPaper_DRA_Weber.pdf
- Williams, E. J. (1999). *Developmental Reading Assessment: Reliability study 1999*. Unpublished manuscript. Retrieved July 23, 2006 from http://www.pearsonlearning.com/correlation/rsp/ResearchPaper_DRA.rtf

August 25, 2006

Natalie Rathvon, Ph.D.
Assistant Clinical Professor, George Washington University, Washington DC
Private Practice Psychologist and School Consultant, Bethesda, MD