

Developmental Reading Assessment

Reliability Study

1999

E. Jane Williams, Ph.D.

In spring of 1999 a reliability study was conducted to examine (1) inter-rater agreement of teachers using the assessment, and (2) internal consistency of the Developmental Reading Assessment (DRA) instrument. Additional data were obtained to examine construct validity as well. Primary teachers from around the country responded to a call requesting participants for this reliability study.

Procedures

A nationally representative sample of students (N = 306) from kindergarten through third grade was included in this study. Each teacher who participated was asked to assess 3 or more children from his or her classroom by conducting and audio taping the DRA conference. After the classroom teacher audiotaped each student's conference, the tape was sent to a second and then a third person to rate. Eighty-seven (87) teachers from 10 states participated as originating teachers. All participants (N=127, originating teachers and raters) had prior experience administering the DRA.

Each originating teacher received a packet, which contained the following:

- ∞ Procedures to follow for the reliability study;
- ∞ A videotape showing examples of children being assessed at 3 different levels;
- ∞ Reporting forms, i.e., Student Reporting Form, Teacher Information Sheet;
- ∞ Audiotapes for checking reliability; and
- ∞ Individual envelopes for returning each student's audiotape and forms.

Original DRA conferences were completed during the last six-weeks of the school year. Returned audiotapes were randomly assigned within grade ranges to blind-raters who were outside the district of the originating teacher. Grade ranges consisted of either kindergarten and first grade, or second and third grade. Each rater was sent appropriate materials (procedures, forms, the DRA continuum, and a return mailing envelope) with the audiotape.

Upon return of the second rating, the tape and new materials were then mailed to a third group of blind-raters. Again each group of blind-raters was randomly assigned conferences originating outside their district. Sixteen (n=16) originating teachers plus 40 new teachers served as second and third raters.

Participants. Participating teachers were from eleven states and 42 districts. Of the 127 teachers, 10 taught kindergarten, 38 taught first grade, 24 taught second grade, 17 taught third grade, and 28 taught other areas, which included reading specialists, learning disability teachers, and literacy coordinators (10 without information). Twenty-one percent of the teachers received 2-days of training within their district, 27 percent received one-day of training, 28 percent received a half-day training, and the remaining received various other formats of training for the DRA. In addition, 73 percent of the teachers had administered the DRA 25 or more times prior to conducting the reliability assessments.

Each originating classroom teacher selected three or more students of varied reading competencies. These students in kindergarten through third grade read assessment texts ranging from level A through level 44 (a total of 20 different text levels ranging from pre-primer to fifth grade). Table 1 shows the number of boys and girls who read at each DRA text level by grade level.

Reliability. To determine the extent of inter-rater agreement among teachers/raters, Rasch rating scale analyses were conducted across 4-facets. Rasch (facet) rating scale analyses were employed to capture the multiple facets of the reading process and their interdependencies. The raters (n=3), students (n=306), text reading levels (n=19), and items (n=5) identified the facets. Items included rating scale responses indicated by teachers for students' rate of accuracy, level of understanding (comprehension), reading stage, phrasing, and reading rate. The rating scale responses to the items are the results of the interactions between the facets.

Cronbach's alpha was employed to determine the internal consistency of the items and text. The item and text separation reliability in Rasch rating scale analyses are equivalent to Cronbach's alpha.

Table 1. Frequency of DRA Text Levels for Children by Grade Level

DRA Text Level	Kindergarten			First Grade			Second Grade			Third Grade			Other	
	N	Boy	Girl	NA	Boy	Girl	NA	Boy	Girl	NA	Boy	Girl	NA	NA
A	4	0	2	0	0	0	2	0	0	0	0	0	0	0
1	9	3	2	0	1	0	3	0	0	0	0	0	0	0
2	9	1	2	0	2	0	4	0	0	0	0	0	0	0
3	17	6	2	0	3	4	1	0	0	0	1	0	0	0
4	14	2	1	0	5	3	2	0	1	0	0	0	0	0
6	13	1	2	0	7	2	0	0	0	0	0	0	0	1
8	9	1	0	0	2	3	2	1	0	0	0	0	0	0
10	7	0	1	0	1	1	1	2	0	0	0	0	0	1
12	10	0	1	0	3	4	0	1	0	0	1	0	0	0
14	12	0	1	0	7	3	0	1	0	0	0	0	0	0
16	20	1	3	0	3	7	0	2	1	0	0	0	0	3
18	23	1	0	0	5	5	0	6	1	0	2	2	0	1
20	14	0	0	0	7	3	0	3	0	0	0	0	0	1
24	24	0	0	0	5	6	0	4	3	0	6	0	0	0
28	30	0	0	0	5	8	0	6	6	0	0	4	0	1
30	17	0	0	0	1	0	0	5	5	0	3	3	0	0
34	8	0	0	0	0	1	0	2	2	0	1	2	0	0
38	39	0	0	0	1	0	0	7	6	0	18	7	0	0
40	15	0	0	0	0	1	0	2	2	0	6	4	0	0
44	12	0	0	0	1	0	0	1	2	0	3	5	0	0
Total	306	16	17	0	59	51	15	43	29	0	41	27	0	8

Teachers reported each participating student’s rate of accuracy, level of understanding (comprehension), reading stage, as well as the teacher’s evaluation of the student’s phrasing and reading rate, which was gathered during the DRA conference and recorded on the DRA Observation Guide.

Rate of accuracy was determined by counting the number of uncorrected miscues and words given by the teacher during the oral reading of a portion or the entire text for the selected DRA text level. The accuracy rates were grouped in the following categories, where: 1 = less than 90 percent accuracy, 2 = 90 through 93 percent accuracy, 3 = 94 through 97 percent accuracy, and 4 = 98 through 100 percent accuracy.

Teachers judged the students’ degree of understanding based on the their retellings using the descriptors on the DRA Continuum. The following four-point scale represents the degree of understanding: (1) little understanding, (2) some understanding, (3) adequate understanding, and (4) very good understanding, Table 2.

Table 2. DRA Levels of Understanding/Comprehension

Retelling represents:			
Little Understanding	Some Understanding	Adequate Understanding	Very Good Understanding
<ul style="list-style-type: none"> ∞ Unorganized ∞ Important details missing ∞ Incorrect information ∞ Misinterpretation 	<ul style="list-style-type: none"> ∞ Somewhat organized ∞ Focuses on parts rather than the whole ∞ Events out of sequence ∞ Includes some details about characters and events ∞ Some misinterpretation 	<ul style="list-style-type: none"> ∞ Organized but may be choppy ∞ Sequential for the most part ∞ Includes main ideas, details about characters, settings, and events ∞ Literal interpretation 	<ul style="list-style-type: none"> ∞ Effectively organized and fluid ∞ Sequential ∞ Includes main idea, important details about characters, settings, and events ∞ Reveals use of background knowledge and experience to interpret story ∞ Uses vocabulary/special phrases from the story

Students' Reading Stages were identified by the teachers based on the level of the text read by the student and other descriptors circled on the DRA Continuum for (1) Emergent, (2) Early, (3) Transitional, and (4) Extending. Generally Emergent readers read DRA text levels A to 2, Early readers read DRA text levels 3 to 10, Transitional readers read DRA text levels 12 to 24, and Extending readers read DRA text levels 28 through 44.

Teachers determined if students' phrasing was appropriate for their reading stage of development based on whether or not the children read word by word, in short phrases, in longer phrases, or observing punctuation as well as if they reread to rephrase or observe punctuation. After deciding the students' levels of phrasing, teachers then determined the extent to which they agreed the levels were appropriate for their reading stages of development. The extent of agreement was based on the following scale, where 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree.

The same scale was used to determine if students' reading rates were adequate for their reading stages of development based on information checked in the DRA Observation Guide. The descriptors for reading rate were slow, inconsistent, adequate, too fast, and adjusted appropriately. The frequency of responses by teachers/raters to the rating scale items is reported in Table 3.

Table 3. Frequency of Responses to Items by Teachers/Raters

Item	N	%
Understanding/Comprehension		
∞ Little understanding	39	4.2
∞ Some understanding	263	28.6
∞ Adequate understanding	334	36.3
∞ Very good understanding	161	17.5
∞ Does Not Apply	124	13.5
Reading Stage		
∞ Emergent	74	8.0
∞ Early	185	20.1
∞ Transitional	357	38.8
∞ Extended	296	32.1
Phrasing is Adequate for Reading Stage		
∞ Strongly Disagree	7	0.8
∞ Disagree	129	14.0
∞ Agree	621	67.4
∞ Strongly Agree	160	17.4
Reading Rate is Adequate for Reading Stage		
∞ Strongly Disagree	13	1.4
∞ Disagree	116	12.6
∞ Agree	616	66.9
∞ Strongly Agree	174	18.9
Information was Helpful with Instruction?		
∞ Strongly Disagree	2	0.2
∞ Disagree	18	2.0
∞ Agree	526	57.2
∞ Strongly Agree	371	40.3

Results

Inter-rater Reliability. Analyses revealed reliability between the originator and second rater was strong, i.e., inter-rater agreement between the first two raters was 0.80, when calculated across facets; inter-rater agreement among all three raters was not as strong. The inter-rater reliability among the three raters was 0.74 across students, text levels and items. Inter-rater agreement was calculated using Rasch rating scale analysis, 4-facet model, by Wright & Stone (1979).

Internal Consistency. The internal consistency was found to be quite strong for the five rating scale items, i.e., item separation reliability (Cronbach's alpha = 0.98), across all three raters as well as for the DRA assessment texts, i.e., text separation reliability (Cronbach's alpha = 0.97).

Construct Validity. The study described was designed to examine the inter-rater reliability and internal consistency of the DRA. Subsequently, additional data were obtained from one school district to help establish the construct validity of the DRA. To ensure that the DRA measured what was intended, the validity of the DRA instructional reading level was assessed. To assess its validity, individual scores on the DRA for the second grade population (N=2470) at the end of the 1998-99 school year from a large urban/suburban school district were correlated with the students' scores from fall of third grade on the Iowa Test of Basic Skills Subscales: Vocabulary, Reading Comprehension, and Total Reading. All correlations were significant at the 0.01 level (2-tailed) using Spearman's Rho rank order correlation; however, the highest and most meaningful correlation for this assessment was with Total Reading ($r = 0.71$, $p < .01$) (see Table 4).

Thus, since the DRA instructional reading levels demonstrated a strong correlation with the Iowa Test of Basic Skills Total Reading subscale for one large urban/suburban school district, this evidence adds strength to the belief that the DRA validly measures a child's ability to decode and understand/comprehend what he/she has read. Further support is provided by the Vermont – DRA validity and reliability report (Biggam, & Grainger, 1998). Biggam and Grainger noted, "A new variation on the notion of content validity is authenticity. Authenticity is analogous to curricular validity in the sense that with both concepts, a test is compared to some external standard of appropriateness...With authenticity, the external standards are various types of literacy tasks that people engage in across a variety of literacy environments...The key question is, Does this test reflect the ways in which we can expect students to use literacy for communication and learning

Table 4. Spearman’s Rho Correlations Between ITBS Subscales and DRA Instructional Levels.

Spearman’s rho Rank Order Correlation	ITBS Reading Comp NCE	ITBS Total Reading NCE	ITBS Vocabulary NCE	DRA Instructional Level
ITBS Reading Comp NCE	----	0.953** 2470	0.812** 2470	0.675** 2470
ITBS Total Reading NCE		----	0.947** 2470	0.708** 2470
ITBS Vocabulary NCE			----	0.675** 2470
DRA Instructional Level				----

** . Correlation is significant at the 0.01 level (2-tailed).

purposes?” (cited by Biggam & Grainger, page 4, from Taylor, Harris, Pearson, & Garcia, 1995, p. 329). The DRA is an authentic performance based assessment in which children are responding to real text through retelling.

In addition the Vermont Department of Education and University of Vermont (with support from the Northeast & Islands Regional Lab at Brown) are currently conducting a further validity study on the VT—DRA (Lipson, M., Biggam, S., Connor, D., & Mekkelsen, J., 1999).

Summary and Discussion

In summary, reliability results from Rasch scale (facet) analyses revealed good to fair reliability between raters. Inter-rater agreement, as measured by rater separation reliability, was good (0.80) between the originating teacher and the second rater, while the inter-rater agreement found among all three raters was a little lower (0.74). Each of the analyses was conducted across 4 facets (raters, students, text levels, and items) with the 5 rating scale items (accuracy, comprehension, stage, phrasing, and reading rate). Additional support for construct validity was obtained from data from Ft. Bend ISD, TX as well as from the VT – DRA statewide assessment report by the Vermont Department of Education, 1998.

One might expect the reliability to be higher between the two raters who were rating the audiotapes, than with the originator teacher, but that was not the case. Nonetheless, it is not

surprising that the agreement among all three raters was slightly lower than that for the first two raters (originating teacher and second rater) considering that the third rater received the audiotape and materials for analysis at the beginning of the next school year in August/September, whereas the second rater received the tape and materials in June/July. The third raters were pressed for time due to the additional demands of beginning a new school year. No teachers/raters received additional training prior to conducting ratings.

It should be noted that a major purpose of the DRA is to help guide instruction. Ninety-eight percent (98%) of the teachers and raters agreed or strongly agreed to the statement that the information gained about the reader during the DRA conference helped them better identify things that the child needed to do or learn next. In a joint position statement entitled, *Learning to Read and Write: Developmentally Appropriate Practices for Young Children* adopted in 1998 by the International Reading Association (IRA) and the National Association for the Education of Young Children (NAEYC), it is stated that

Throughout these critical years accurate assessment of children's knowledge, skills, and dispositions in reading and writing will help teachers better match instruction with how and what children are learning. However, early reading and writing cannot be measured as a set of narrowly defined skills on standardized tests. These measures often are not reliable or valid indicators of what children can do in typical practice, nor are they sensitive to language variation, culture, or the experience of young children (Shepard & Smith, 1998; Shepard, 1994; Johnston, 1997). Rather, a sound assessment should be anchored in real-life writing and reading tasks...p.15 and should support "individualized diagnosis needed to help young children continue to progress in reading and writing." p.20

Recommendations

The following steps are recommended to further strengthen the DRA's reliability as a primary reading assessment.

1. Teachers are trained for reliability purposes using audio and videotapes. Reliability training is critical to assure that all students are evaluated based on the same criteria and the score for one child is the same as that for another child with the same responses/behaviors even when evaluated by a different teacher or at a different time.
2. To increase reliability when conducting the DRA, teachers audiotape the assessment conference as suggested by Clay in *An Observation Survey of Early Literacy Achievement* (1993, page 28).

References

- A joint position statement of the International Reading Association (IRA) and the National Association for the Education of Young Children (NAEYC). (1999). *Learning to Read and Write: Developmentally Appropriate Practices for Young Children*. 20.
- Biggam, S., & Grainger, E. (Personal communication, March 1999). Summary of finding from the 1998 statewide administration of the "VT-DRA." Vermont Department of Education.
- Clay, M.M. (1993). *An observation survey of early literacy achievement*. Auckland, New Zealand: Heinemann.
- Criswell, G., & Duvall, C. (November 1999). Electronic data file. Ft. Bend Independent School District, Testing and Evaluation.
- Johnston, (1997). *Knowing literacy: Constructive literacy assessment*. York, ME: Stenhouse.
- Linacre, J.M. (1997). *A user's guide to Facets: Rasch measurement computer program*. Chicago, IL: MESA Press.
- Lipson, M., Biggam, S., Connor, D., & Mekkelsen, J. (1999, December). *Large scale early reading assessment: Challenges, strategies, and implications*. Paper presented at the National Reading Conference, Orlando, FL.
- Shepard, L. & Smith, M.L. 1998. Escalating academic demand in kindergarten: Some nonsolutions. *Elementary School Journal* 89: 135-46.
- Shepard, (1994). The challenges of assessing young children appropriately. *Phi Delta Kappan* 76: 206-13.
- Wright, B.D., and Masters, G.N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.