Technical Review Committee Confirms Highest NCRTI Ratings for Observation Survey of Early Literacy Achievement

Jerome D'Agostino, Director, International Data Evaluation Center

Established by the American Institutes for Research, Vanderbilt University, and the University of Kansas through the U.S. Department of Education's Office of Special Education, the National Center for Response to Intervention (NCRTI) is charged with providing technical assistance to states and districts to implement proven response to intervention (RTI) models.

In May 2011, *An Observation Survey of Early Literacy Achievement* (Clay, 2002, 2005) was submitted to the NCRTI for approval as a screening tool to identify students for RTI, by RRCNA Executive Director Jady Johnson; Billie Askew, Patricia Kelly, and Robert Schwartz, Reading Recovery university trainers; and Jerome D'Agostino, director of the International Data Evaluation Center (IDEC).

In March 2012, after completion of the rigorous two-stage review process, the Observation Survey has received the highest possible ratings on all five of the NCRTI's technical standards: classification accuracy; generalizability; reliability; validity; and disaggregated reliability, validity, and accuracy for subgroups. As a result, the Observation Survey is one of only three reading assessments that received the NCRTI's highest ratings in all categories. Now that the Observation Survey has been approved by the Technical Review Committee of NCRTI, it can be used by school psychologists, special educators, and others as an evidenced-based screening instrument to identify children at risk for literacy failure and thus, likely to need intervention services in reading and writing.

Total Score Scale and Item Response Theory

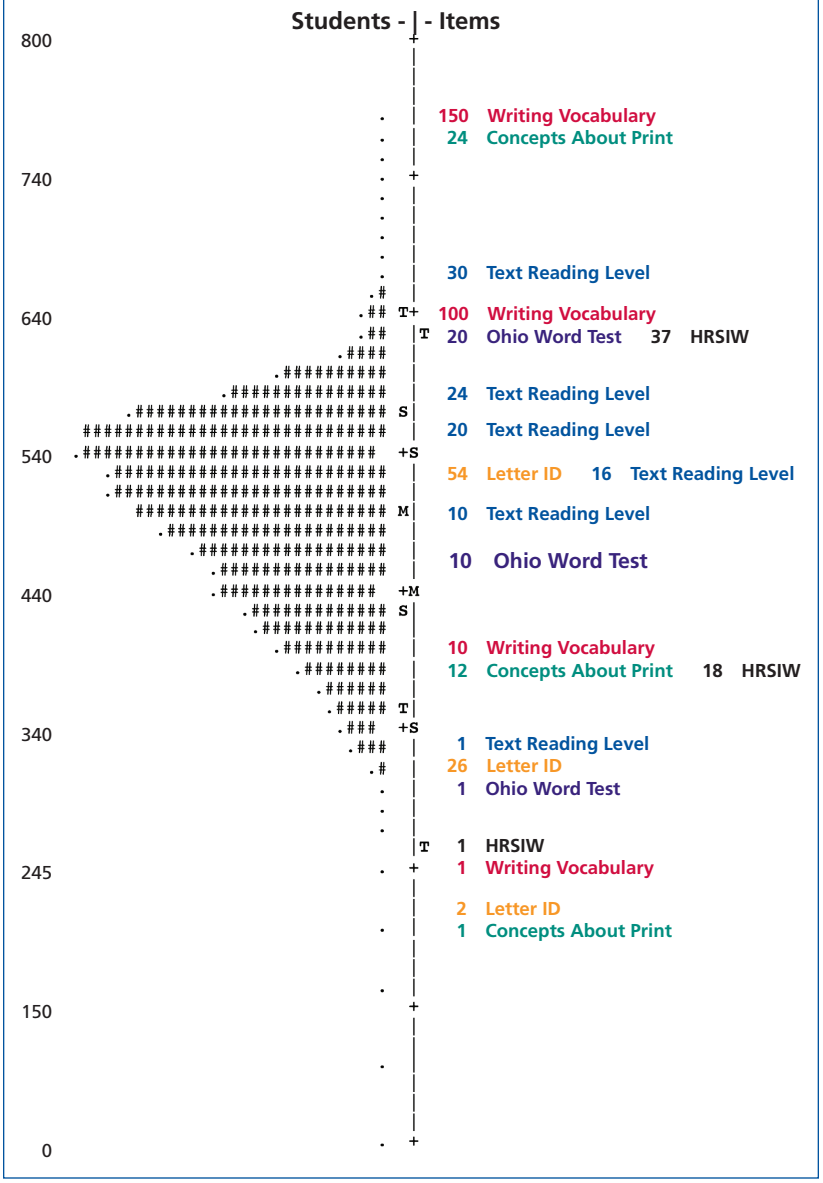
In order to receive the highest rating in all five technical standards, an Observation Survey total score was developed based on students' scores from all six Observation Survey tasks: Letter Identification, Ohio Word Test, Hearing and Recording Sounds in Words, Concepts About Print, Writing Vocabulary, and Text Reading Level. The Observation Survey raw scores of students in the random sample (n = 9,760) collected by IDEC in 2009–2010 were used as a first step. Then a one-parameter item response theory (IRT) measurement model was employed to create the Observation Survey total score scale that can gauge literacy achievement at any point during the school year, and thus, measure change over time. The logic of IRT is to estimate the difficulty of test items or each additional point for partial credit items, and then estimate student proficiency by examining how the student responded to items with varying levels of difficulty. Because the goal was to develop a growth scale, it was critical that children's scores from throughout the year be used to calibrate the item and point difficulties. Using fall, mid-year, and year-end scores of each student would have violated the assumption of score independence, so instead, only one of the three scores of each student was chosen at random to create a scale calibration sample. Thus, about one-third of the sample's fall Observation Survey scores, another one-third of the mid-year scores, and one-third of the year-end scores were used for this scale.

> The Observation Survey can be used by school psychologists, special educators, and others as an evidenced-based screening instrument to identify children at risk for literacy failure.

Students' raw scores on all six Observation Survey tasks from fall, mid-year, or year-end were treated as six partial-credit items in the IRT analysis (e.g., a student who identified 30 letters out of 54 on Letter Identification earned 30 of the possible 54 points). The one-parameter IRT model fit the data well, indicating that the six tasks could be used together to develop one common

literacy scale. (Please note that the total score resulting from IRT procedures will be computed only for research purposes. It will not be an appropriate score, or procedure, for anyone using the Observation Survey with individual children.) The scale is reported initially in logits (log odd units) that vary from about -4 to +4. A linear transformation then was imposed on the logits to create a scale that varies from 0 to 800 points. Figure 1 presents the resulting IRT scale, with the student score distribution on the left and the six Observation Survey task items on the right. The number in front of the task title is the partial credit or raw score on each task (not all points are presented). Items with accompanying points at the top of the scale were harder for students to accomplish. A score of 150 on the Writing Vocabulary task represented the greatest level of literacy achievement produced by the sample children, followed by a perfect 24 on Concepts About Print. The easiest, or most fundamental, score is 1 on Letter Identification, followed by 1 on Concepts About Print, and a score of 2 on Letter Identification. As can be seen, the student score distribution has a slight negative skew, but the items and student scores are spread equally across the scale, indicating the appropriateness of the Observation Survey for measuring literacy development throughout first grade.

Figure 1. Item Map of Observation Survey Total Score



Statistical Results and Evidence Ratings

The NCRTI approval process requires the submission of statistical results and other information that address each of the five technical standards (see NCRTI, 2012). Table 1 presents the criteria required to obtain a top score in each dimension. To receive a *convincing evidence* score for classification accuracy, the measure must have a greater than .85 receiver operator characteristic (ROC) value when predicting an external outcome that meaningfully defines year-end risk of academic failure. Random sample spring Slosson scores were combined across multiple years, and a grade equivalent score of 1.9 (first grade, ninth month) was used to define the cut score between at risk (less than 1.9) and not at risk (1.9 or greater). Students who scored below 1.9 were considered below grade level in spring or end of first grade. Twenty-six percent of the random sample (excluding children who received Reading Recovery) had a spring Slosson grade equivalent less than 1.9. The area under the curve was .87 when fall total scores were used to predict risk or no risk in the spring

based on 1,826 students from the 2009–2010 and 2010–2011 random sample who did not receive Reading Recovery and whose data were available.

For a screening device to receive a rating of *broad* generalizability, the sample must be large, representative, and nationally based. The results also must be cross-validated on another data set. Combining random sample data across multiple years and many states for those students with fall Observation Survey and spring Slosson scores yielded a large and representative national sample. In order to cross-validate, the ROC analysis was performed on random sample data from 2007–2008 and 2008–2009 combined. The area under the curve was .85 for the cross-validation sample of 1,594 children. The alpha coefficient was .87, and the split-half was .89 for 2009–2010 random sample total scores (n = 7,926), which was deemed convincing evidence for the reliability standard.

Content, construct, and predictive validity evidence was produced to fulfill the validity standard. Information about how each of the six tasks measures critical aspects of literacy development and how each task aligns with the national reading standards was submitted to the Technical Review Committee to meet the content validation requirement. Predictive validation involves correlating the measure with other measures administered after a period of time has elapsed. Several correlation coefficients between fall Observation Survey total scores and subsequent measures for the 2009–2010 random sample were provided, including spring Slosson scores (.72), mid-year Slosson scores (.75) spring text reading levels (.74), mid-year total scores (.83), and spring total scores (.73). All values were greater than the .70 required by the NCRTI for its review. There are various forms of construct validity evidence, including correlating the measure with external measures administered at the same time. The correlation between fall total scores and fall Slosson scores for 2009–2010 random sample students was .78, which was sufficient evidence for construct validity. The validity evidence for the Observation Survey was judged to be convincing by the Technical Review Committee.

Table 1. Highest Rating Criteria for the NCRTI Screening Standards and Evidence Provided to Meet Standards

Technical Standards	Highest Rating Criteria	Evidence for OS Total Score
Classification Accuracy	Area under ROC curve must be greater than .85 for screener predicting a criterion	Area under curve was .87 for fall OS total score predicting spring Slosson scores
Generalizability	Large representative national sample with cross validation	Multiple years of random sample data
Reliability	Two or more of split-half, alpha, test-retest, or inter-rater > .80	Alpha coefficient .87; split half .89
Validity	Must show content, construct (> .70), and predictive (> .70) validity	Correlations greater than .70 with various measures; OS aligned with national reading standards and reflects key literacy aspects
Disaggregated Reliability, Validity, Classification	At least two of three (reliability, validity, classification accuracy) for at least one group and meet above criteria	Met above reliability, validity, and classification accuracy for African American and Hispanic students

Technical Standards from NCRTI (2012)

The final technical standard relates to meeting at least two of three (reliability, validity, or classification accuracy) standards for defined subgroups of students. Random sample data from 2009–2010 for African American and Hispanic students were used for the analyses to address the disaggregation standard. For both subgroups separately, sufficient evidence for all three standards was produced to reach the convincing evidence rating.

All of the analyses were conducted on each Observation Survey task separately, but sufficient evidence could not be generated to earn the highest ratings on all five standards for any one task. It is not the case that the tasks alone do not provide critical information about student literacy learning. The major challenge with most Observation Survey

National Center on Response to Intervention
www.rti4success.org

IDEAs that Work
U.S. Office of Special Education Programs

Screening Tools Chart Continued

Tools	Area	Classification Accuracy	Generalizability	Reliability	Validity	Disaggregated Reliability, Validity, and Classification Data for Diverse Populations	Efficiency			
							Administration	Administration & Scoring Time	Scoring Key	Benchmarks/ Norms
Observation Survey of Early Literacy Achievement	*Reading	●	Broad	●	●	●	Individual	15-45 Minutes	Yes	Yes

Legend ● Convincing evidence ◐ Partially convincing evidence ○ Unconvincing evidence — Data unavailable or inadequate
* Added in the 2011 review ! Information updated during the 2011 review

The NCRTI publishes the Screening Tools Chart to assist educators and families in becoming informed consumers who can select screening tools that best meet their individual needs. The chart reflects the results of annual reviews of screening tools by the NCRTI Technical Review Committtee. The Observation Survey appears on page 6 of the chart; portions shown here.

tasks analyzed separately was the highly skewed distribution of the fall scores, such as Letter Identification (most non-Reading Recovery students know most or all letters by fall of first grade) and Text Reading Level (75% of random sample students score 6 or less in fall of first grade), which delimits these tasks from adequately predicting spring outcomes. However, treating the tasks as six partial credit items on a total scale led to fall and spring distributions that were more normally shaped, and this greater score dispersion allowed for better estimates of the relationships with the external measure scores.

Valid Measures and Observational Data to Inform RTI Identification

Even with those measurement reasons, considering all the task scores of the Observation Survey provides for a more-comprehensive indicator of children's literacy levels. As Clay (2002, 2005) pointed out,

> When important decisions are to be made we should increase the range of observations we make in order to decrease the risk that we will make errors in our interpretations. … It is also why a wide range of measures or observations should be made. (p. 12)

The information provided from all six Observation Survey tasks yields more-reliable estimates of literacy, and it covers more thoroughly the domain of the construct to be measured. The total scores, however, are not meant to supplant knowledgeable teachers' skillful analysis and interpretation of students' performance on the Observation Survey. Teachers must consider the *entire* profile of student scores on all six tasks and evaluate the full ensemble of each child's reading and writing behaviors to inform instruction..

Yet certain identification processes, such as with many RTI models, require the use of a single test score to screen students. The evidence submitted to NCRTI and judged by the Technical Review Committee to fully meet their standards for screening devices indicates that the Observation Survey task scores taken together are very useful to fulfill that purpose with young children.

References

Clay, M. M. (2002, 2005). *An observation survey of early literacy achievement* (2nd ed., rev. 2nd ed.). Portsmouth, NH: Heinemann.

National Center on Response to Intervention. (2012). Screening Charts. Retrieved April 3, 2012, from http://www.rti4success.org/screeningTools.

About the Author

Jerome D'Agostino is a professor in the Quantitative Research, Evaluation, and Measurement program at The Ohio State University and director of the International Data Evaluation Center. He specializes in assessment, measurement, and intervention evaluation. Dr. D'Agostino also is director of the i3 project to scale up Reading Recovery.