# Report from the International Data Evaluation Center: 25 Years of Consistently High Results

*Jerome D'Agostino, International Data Evaluation Center*
*Anne-Evan Williams, The Ohio State University*

Over its 25-year history in the United States, Reading Recovery has been committed to collecting and analyzing data on students and teachers to examine the effectiveness of the intervention. This report will describe general summary information about Reading Recovery and Descubriendo la Lectura (DLL) participants during the 2009–2010 school year, and present literacy outcomes analyses of Reading Recovery and DLL students over the school year compared to both a random sample and low random sample of nonparticipants. The 2009–2010 school year was no different from many past years — Reading Recovery and DLL participants significantly outperformed the comparable comparison group and significantly closed the achievement gap with random sample students in general.

*At the end of the school year, these Reading Recovery students celebrated their ability to R-E-A-D. Results of all first graders with Reading Recovery lessons are reported by the International Data Evalution Center at The Ohio State University. No other reading intervention has the history and ongoing capacity to measure results for each student taught.*

Reading Recovery students whose lessons were discontinued had an average end-of-year text reading level of 19.5 in 2009–2010, compared to 19.2 the prior year. The mean text reading level values for DLL students were 19.0 in 2008–2009 and 18.7 in 2009–2010. These differences are minor and could have resulted from random fluctuation across years. The consistency of the scores indicates the stability of the interventions to produce positive results for students.

Unlike past annual evaluations that relied on text reading level as the primary outcome, however, we computed a composite score for each child based on scores of all six Observation Survey subtests. Although text reading captures the essence of the reading process perhaps better than any of the six survey measures, it suffers from a floor effect in fall among Reading Recovery and similarly initially low-scoring students. The composite measure, however, does not have this problem because it includes Observation Survey subtests that do not have a fall floor effect. Thus, the composite measure is superior for capturing student growth over the full span of first grade, and thus, is more appropriate for evaluating the effects of Reading Recovery and DLL. This is a practice appropriate for the statistical analyses conducted to address our research questions.

## Summary of Reading Recovery Outcomes

Table 1 provides the overall participation numbers of students, teachers, training sites and states involved in Reading Recovery and Reading Recovery data collection through IDEC. As can be seen, 73,248 Reading Recovery children were served by 8,785 Reading Recovery teachers in 5,412 schools in 1,721 school districts. The teachers were supported by 411 teacher leaders at 328 teacher trainer sites across 44 states and federal entities. Twenty-one university training centers provided professional development and support. At each participating school, two children chosen at random comprised the random sample and were asked to take the Observation Survey at three time points during the year (fall, mid-year, and year-end). Some of these also received Reading Recovery during the year. There were over 10,000 children who served as a random sample, which includes some Reading Recovery students.

**Table 1. Participation in Reading Recovery in the United States 2009–2010**

| Entity | n |
|---|---|
| University Training Centers | 21 |
| Teacher Training Sites | 328 |
| States and Federal Entities* | 44 |
| Systems | 1,721 |
| Buildings | 5,412 |
| Teacher Leaders | 411 |
| Teachers | 8,785 |
| Reading Recovery Students | 73,248 |
| Random Sample for RR | 10,160 |

*including Bureau of Indian Affairs, Department of Defense Domestic, and Department of Defense Overseas

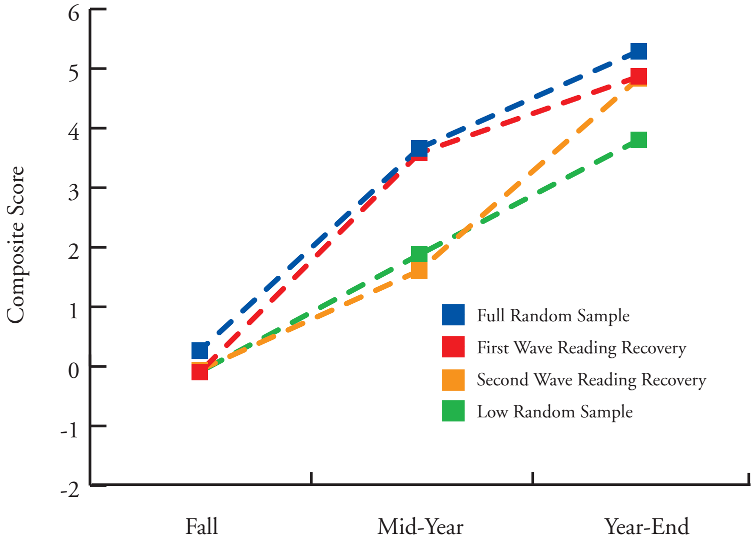
Students from diverse backgrounds throughout the country participated in Reading Recovery during the 2009–2010 year. Fifty-eight percent were boys and 63% were eligible for free or reduced-price lunch. Children were from many ethnic backgrounds: 59% White; 18% African American; 17% Hispanic; 2% Asian American; 1% Native American; and 3% represented multiple races or other ethnic backgrounds. Among all participants, 60% reached the average level of performance of their peers and their lessons were discontinued. Another 20% were recommended for further evaluation, while 14% received incomplete interventions, 4% moved during instruction, and 2% were classified as none of the above. Of the children served who received a complete intervention, 75% reached average levels of performance among their peers in a mean of 15.5 weeks.

To examine the achievement performance of Reading Recovery students, we considered two distinct comparison groups that address particular research questions. To answer the question of what extent Reading Recovery students close the achievement gap with their peers in general, we analyzed the differences of Reading Recovery students and the random sample (with Reading Recovery students removed) on a composite measure of the Observation Survey. To address the question of what would have been the expected growth from the beginning to the end of the school year for Reading Recovery students had they not received the intervention, we examined the difference between Reading Recovery students and the bottom 25% of students from the random sample (with Reading Recovery students not included) on the composite indicator. We selected the lowest 25% of random sample students for the second research question because that group had the closest mean composite values to Reading Recovery students on the initial administration of the Observation Survey.

We created the composite scores by using principal component analysis on the scores from the six survey tests. This method produces a weighted linear composite of the six tests (in a z-score metric), where the weight for each test is based on how well the test is intercorrelated with the other tests that comprise the composite. We computed the weights based on the fall random sample data, including Reading Recovery students who were part of the random sample. Then we computed all Reading Recovery students' fall composite scores based on the weighted linear model generated from the analysis. To reflect growth from fall to mid-year and year-end, we converted the mid-year and year-end scores to z-scores based on the fall means and standard deviations before computing the composite values. This system led to standardized composite scores for each student at three time points during the school year that reflected growth based on fall values.

Figure 1 presents the mean composite values for Reading Recovery students whose lessons were successfully discontinued and who were served in the first round or wave of the school year, Reading Recovery students served second (Wave 2), random sample, and low random sample students. As can be seen from the figure, the low random sample group had comparable fall composite scores,

**Figure 1.** **Mean 2009-2010 Observation Survey Composite Values for Reading Recovery Students with Discontinued Status, Full Random Sample Students, and Low Random Sample Students**



on average, compared to the Reading Recovery students (especially Wave 1 students). By mid-year, Wave 1 students had a significantly greater mean gain relative to both untreated low groups (the Wave 2 and low random sample students), and had a comparable mean score to the full random sample. The latter finding was anticipated because students' lessons were discontinued when they reached average performance levels. By year-end, Wave 1 students dropped off to some degree from the average random sample gain, but retained a significant average gap from low random sample students. The mean Wave 2 student reached the average of Wave 1 students. Note also that Wave 2 students, on average, actually had significantly smaller fall-to-mid-year gains than low random sample students—perhaps indicating accurate identification of children in need of Reading Recovery by teachers—and that the low random sample represented an adequate scenario of performance of struggling students without the intervention.

We also examined the magnitude of mean differences, or effect sizes, between Reading Recovery students and low and full random sample students. Tables 2 and 3 present the mean composite and Observation Survey scores of Waves 1 and 2 Reading Recovery students combined and full random sample and low random sample students respectively. In both tables, the right-hand columns provide the effect sizes in terms of standardized mean differences (positive values indicate that the Reading Recovery mean was greater than the comparison mean value) and the percentile standing of the average Reading Recovery child in the comparison group distribution (in parentheses). As expected, the mean Reading Recovery scores in fall ranged from the 14th to 29th percentile, with the latter value likely due to an apparent ceiling effect of Letter

**Table 2.** **Comparison of 2009-2010 Fall and Year-End Mean Scores with Effect Sizes on the Tasks of the Observation Survey for Reading Recovery Discontinued and Full Random Sample Students**

| | Discontinued<br>(n = 28,335) | | Random Sample<br>(n = 6,930) | | Effect Size<br>Difference | |
|---|---|---|---|---|---|---|
| Observation Survey Task | Fall | Year-End | Fall | Year-End | Fall | Year-End |
| Composite Score | -.84 | 4.90 | 0.27 | 5.29 | -1.10 (14) | -.25 (40) |
| Text Reading Level | 1.30 | 19.50 | 6.20 | 21.80 | -.81 (21) | -.34 (37) |
| Writing Vocabulary | 11.10 | 56.30 | 22.10 | 58.30 | -.90 (19) | -.11 (46) |
| Hearing and Recording Sounds in Words | 20.50 | 36.00 | 29.40 | 35.90 | -1.01 (16) | +.03 (49) |
| Letter Identification | 49.00 | 53.50 | 51.80 | 53.60 | -.57 (29) | -.04 (48) |
| Ohio Word Test | 3.60 | 19.20 | 10.20 | 19.20 | -1.07 (14) | .00 (50) |
| Concepts About Print | 12.70 | 21.00 | 15.80 | 21.00 | -.87 (19) | .00 (50) |

Table 3. **Comparison of 2009-2010 Fall and Year-End Mean Scores with Effect Sizes on the Tasks of the Observation Survey for Reading Recovery Discontinued and Low Random Sample Students**

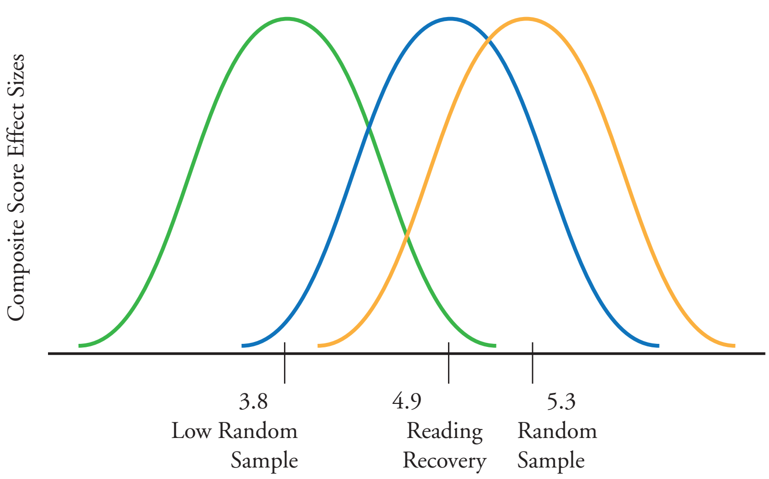
| | Discontinued<br>(n = 28,335) | | Random Sample<br>(n = 1,596) | | Effect Size<br>Difference | |
|---|---|---|---|---|---|---|
| Observation Survey Task | Fall | Year-End | Fall | Year-End | Fall | Year-End |
| Composite Score | -.84 | 4.9 | -.91 | 3.8 | +.06 (52) | +.66 (75) |
| Text Reading Level | 1.30 | 19.5 | 1.60 | 15.6 | -.04 (49) | +.57 (72) |
| Writing Vocabulary | 11.10 | 56.3 | 9.90 | 46.9 | +.09 (54) | +.51 (69) |
| Hearing and Recording Sounds in Words | 20.50 | 36.0 | 19.30 | 34.2 | +.14 (55) | +.63 (74) |
| Letter Identification | 49.00 | 53.5 | 48.70 | 53.1 | +.05 (52) | +.25 (60) |
| Ohio Word Test | 3.60 | 19.2 | 3.50 | 17.7 | .00 (50) | +.61 (72) |
| Concepts About Print | 12.70 | 21.0 | 12.40 | 19.2 | +.80 (53) | +.71 (76) |

Identification in the random sample. By year-end, the effect size differences shrink significantly, indicating the closing of the achievement gap. On the composite measure, Reading Recovery students, on average, performed at the 40th percentile relative to the random sample. On Hearing and Recording Sounds in Words, the mean Reading Recovery score actually was larger; and on the Word Test, Letter Identification, and Concepts About Print, the average Reading Recovery score was at the average of the comparison group. Ceiling effects on those measures, however, may have contributed to the equivalent mean values. On the two measures with no ceiling, Vocabulary and Text Reading Level, the average Reading Recovery students were at the 46th and 37th percentile, respectively.

Table 3 reveals that Wave 1 and Wave 2 Reading Recovery students whose lessons were discontinued had slightly larger mean fall scores than low random sample students on five of the seven indicators, but had slightly lower Text Reading level average scores and comparable Word Test values. The larger positive effects at year-end indicate the significant gain increase of Reading Recovery students relative to the comparison group. On the composite, the average Reading Recovery student performed at the 75th percentile of the low random sample, indicative of a large intervention effect.

Figure 2 provides a visual interpretation of the year-end effect sizes on the composite measure. On the horizontal axis are the mean scores of the low random sample, Reading Recovery students, and random sample. Around each mean is a hypothetical bell curve indicating the expected distribution of the scores in each population among the three groups. The Reading Recovery mean of 4.9 (which is at the 50th percentile in the Reading Recovery distribution) would be at the 75th percentile in the low random sample distribution, and at the 40th percentile in the random sample distribution. Note that the Reading Recovery

Figure 2. **2009-2010 Year-End Composite Score Effect Sizes Comparing the Reading Recovery, Low Random Sample, and Random Sample Means**



student distribution is closer to the random sample than the low random sample by the end of the year.

## Summary of Descubriendo la Lectura Outcomes

Descubriendo la Lectura (DLL), the reconstruction of Reading Recovery in Spanish, is for first graders who receive their initial literacy instruction in Spanish. During the 2009–2010 school year, 803 DLL children were taught by 110 teachers (see Table 4). These children were served in 95 schools in 27 school districts in seven states. The teachers received professional development support from 31 teacher leaders. Of the children who participated in DLL, 60% were boys and 97% were Hispanic (2% were multiple ethnicity students and 1% were White, not Hispanic). Ninety-five percent of DLL participants qualified for free or reduced-price lunch.

Among all children served in DLL, 55% reached the average reading levels of their peers, and their lessons were discontinued within an average time of 15.3 weeks. Another 22% were recommended for further evaluation, 4% moved, and 16% received incomplete interventions.

**Figure 3. Mean 2009-2010 Instrumento de Observacíon Composite Values for Descubriendo la Lectura Students with Discontinued Status, Full Random Sample Students, and Low Random Sample Students**

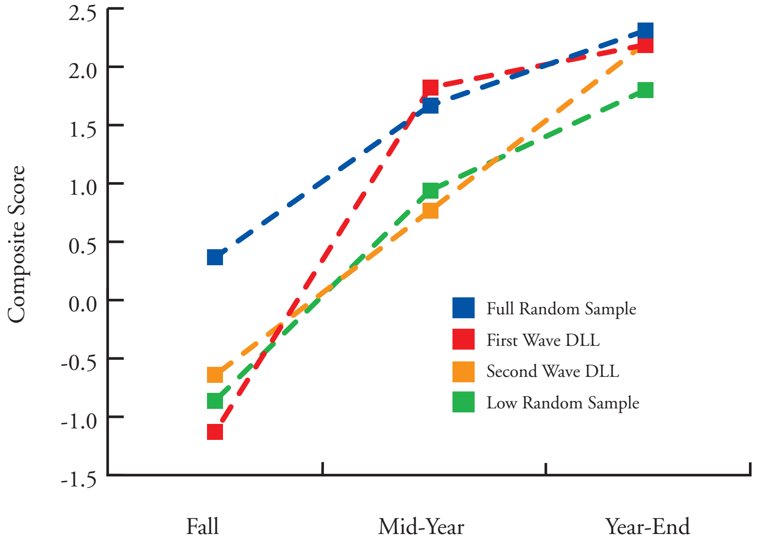


**Table 4. Participation in Descubriendo la Lectura in the United States 2009–2010**

| Entity | n |
|---|---|
| University Training Centers | 5 |
| Teacher Training Sites | 26 |
| States | 7 |
| Systems | 27 |
| Buildings | 95 |
| Teacher Leaders | 31 |
| Teachers | 110 |
| DLL Students | 803 |
| Random Sample for DLL | 168 |

Following the same sampling processes as Reading Recovery, two students per participating DLL school were administered the Instrumento de Observación (the Spanish version on the Observation Survey) in fall, mid-year, and at the end of year. We followed the same procedures as we did for the Reading Recovery data to identify a low random sample group, compute composite scores at each time point, and compare mean differences. Figure 3 presents the mean scores for both Wave 1 and 2 students whose lessons were successfully discontinued and all DLL random sample participants on the composite at each time point. Tables 5 and 6 provide the mean scores for all DLL students combined, and full and low random sample students in fall and at the end of year. The effects sizes are included in both tables.

The graph indicates that the results were quite similar to those for Reading Recovery with a few exceptions. As is evident from Figure 3 and Table 5, DLL students had considerably lower composite scores than the full random sample at the beginning of the year. DLL students were at least one standard deviation unit lower than random sample students on all measures of the survey and the composite, which placed DLL students at the 9th to 16th percentile of their peers.

By mid-year, however, Wave 1 DLL students actually surpassed full random sample students, but then fell behind them—on average—at the end of the year. Wave 2 students gained less from fall to mid-year relative to the low random sample students, but then surpassed them during their treatment period. Both DLL waves finished the year with similar average composite values. Combined, the average DLL student

**Table 5. Comparison of 2009-2010 Fall and Year-End Mean Scores with Effect Sizes on the Tasks of the Instrumento de Observacíon for Descubriendo la Lectura Discontinued and Full Random Sample Students**

| | Discontinued<br>(n = 275) | | Random Sample<br>(n = 104) | | Effect Size<br>Difference | |
|---|---|---|---|---|---|---|
| Instrumento de Observacíon Task | Fall | Year-End | Fall | Year-End | Fall | Year-End |
| Composite Score | -.99 | 2.19 | .37 | 2.31 | -1.36 (9) | -.17 (43) |
| Análisis Actual del Texto | .80 | 18.70 | 4.90 | 21.20 | -1.00 (16) | -.39 (35) |
| Escritura de Vocabulario | 9.00 | 46.40 | 19.40 | 45.60 | -1.07 (14) | +.05 (52) |
| Oír y Anotar los Sonidos en las Palabras | 22.00 | 38.30 | 33.80 | 38.50 | -1.14 (13) | -.09 (47) |
| Identificacíon de Letras | 45.10 | 58.70 | 55.40 | 59.40 | -1.06 (15) | -.24 (41) |
| Prueba de Palabras | 5.90 | 19.60 | 14.30 | 19.50 | -1.33 (9) | +.04 (52) |
| Conceptos del Texto Impreso | 10.20 | 19.80 | 14.30 | 19.90 | -1.02 (15) | -.03 (49) |

performed at the 43rd percentile of random sample students (see Table 5) and the 65th percentile of their low random sample peers (Table 6) on the composite measure by year-end. On some survey tests—such as the word (Prueba de Palabras) and vocabulary (Escritura de Vocabulario) tests—the DLL means values were significantly larger than the random sample averages (see Table 5). Yet on text reading (Análisis Actual del Texto), DLL students were at the 35th percentile of random sample students. Because Análisis Actual del Texto scores at year-end varied considerably, student's text reading scores factored heavily in their composite values.

Only 18 students comprised the low sample, which resulted in larger standard errors around the means for that group. Even with this small sample size, nonetheless, the noted mean differences were statistically significant. The DLL fall means values were slightly lower overall than the low random sample, but were higher on two of the six Instrumento de Observacíon tests. By year-end, the DLL average scores across tests were at the 51st to 79th percentile of the comparison group (see Table 6). The small effect found on the Oír y Anotar los Sonidos en las Palabras, which is the Spanish version of the Hearing and Recording Sounds in Words test, likely was due to a ceiling effect (the maximum possible score was 39 and both group means were 38.3). The 0.39 effect size on

**Table 6. Comparison of 2009-2010 Fall and Year-End Mean Scores with Effect Sizes on the Tasks of the Instrumento de Observacíon for Descubriendo la Lectura Discontinued and Low Random Sample Students**

| | Discontinued<br>(n = 275) | | Random Sample<br>(n = 104) | | Effect Size<br>Difference | |
|---|---|---|---|---|---|---|
| Instrumento de Observacíon Task | Fall | Year-End | Fall | Year-End | Fall | Year-End |
| Composite Score | -.99 | 2.19 | -.87 | 1.8 | -.13 (45) | +.39 (65) |
| Análisis Actual del Texto | .80 | 18.70 | 1.10 | 17.8 | -.05 (48) | +.14 (56) |
| Escritura de Vocabulario | 9.00 | 46.40 | 8.70 | 39.0 | +.03 (51) | +.47 (68) |
| Oír y Anotar los Sonidos en las Palabras | 22.00 | 38.30 | 22.80 | 38.3 | -.07 (47) | +.03 (51) |
| Identificacíon de Letras | 45.10 | 58.70 | 49.60 | 57.8 | -.47 (32) | +.36 (64) |
| Prueba de Palabras | 5.90 | 19.60 | 7.10 | 18.5 | -.19 (43) | +.47 (68) |
| Conceptos del Texto Impreso | 10.20 | 19.80 | 9.60 | 16.9 | +.14 (56) | +.80 (79) |

the year-end composite scores reflects a moderate intervention effect, but the reliability of this effect estimate was compromised given the small sample size of the comparison group.

Note that the gains on the composite measure appear to be much greater for Reading Recovery than for DLL (see Figures 1 and 3). Reading Recovery students gained about six standard deviation units from fall to year-end, whereas DLL students gained about three standard deviations over the same period. The English and Spanish survey scales, however, are not directly comparable because the tests are not exactly the same, and the composite values are based on the respective random sample student performance in each subsample. The most likely reason for the scale difference is the larger gains in vocabulary among Reading Recovery and random sample students compared to DLL and DLL random sample students. Reading Recovery and DLL students' text reading levels at the end of the year, however, were comparable.

*Over its 25-year history in the United States, Reading Recovery has been committed to collecting and analyzing data on students and teachers to examine the effectiveness of the intervention. The 2009–2010 school year was no different from many past years — Reading Recovery and Descubriendo la Lectura participants significantly outperformed the comparable comparison group and significantly closed the achievement gap with random sample students in general.*

## Conclusion

The major focus of the national evaluation is to ascertain if Reading Recovery and Descubriendo la Lectura provide the lowest-achieving first-grade children with the high-quality interventions they need to close the literacy gap with their peers. Two groups of peers are defined to address two distinct questions about the effectiveness of the interventions: To what extent do Reading Recovery and DLL students reach national average levels of achievement, and what would be the literacy achievement trajectory if eligible students did not receive the interventions?

To answer the first research question, two students were chosen at random from each Reading Recovery or DLL school to represent a random sample. We removed intervention students from those groups to compare treated and untreated students' achievement levels. It is critical to note that comparison students in this case did not exactly represent the "national distribution" of Observation Survey or Instrumento de Observacíon scores. In order to produce a sample representative of students nationally, the entire population of first-grade students in the country would need to be considered, including students at schools without Reading Recovery or DLL. IDEC, however, does not sample from schools without the interventions. Furthermore, a sample representative of the national first-grade population would have to include Reading Recovery or DLL

> Perhaps like no other intervention, Reading Recovery has embraced evaluation since its inception. The results reveal the year-to-year consistency of Reading Recovery in terms of providing struggling first-grade students the opportunity to get back on track toward academic success.

students, which would involve ascertaining the proportion of students in the country who receive the interventions. The sampling methods used for the national evaluation do not meet these requirements. But given that we removed Reading Recovery or DLL students from the respective comparison groups, we likely produced conservative intervention effects because their inclusion would have lowered the average year-end random sample scores overall.

Presently, IDEC is considering a sampling method that would include waiting list students from Reading Recovery or DLL schools. Though the students in this group would be similar to low random sample students in some regards—such as being initially more proficient compared to students provided the intervention—they will come from the same schools, which will allow us to compute school-level effects that can be aggregated to the national level. Statistically adjusting for fall scores or relying on another method known as *regression discontinuity* will allow us to render more-accurate intervention effects.

Perhaps like no other intervention, Reading Recovery has embraced evaluation since its inception and has relied on annual results to support its continuation. The results in its 25th year in Unites States schools reveal the year-to-year consistency of Reading Recovery in terms of providing struggling first-grade students the opportunity to get back on track toward academic success.

## About the Authors

Dr. Jerome D'Agostino is an associate professor in the Quantitative Research, Evaluation, and Measurement program at The Ohio State University. He specializes in assessment, measurement, and intervention evaluation. He received his PhD in Measurement, Evaluation, & Statistical Analysis from The University of Chicago in 1997.

Anne-Evan Williams is a doctoral student in the Quantitative Research, Evaluation, and Measurement program at The Ohio State University, and has worked in the publishing industry in literacy assessment development. She holds a masters degree in Education in Advanced Reading and Language Leadership from the University of California, Berkeley, and taught as a K–12 reading specialist for 8 years.