R2LP's Early Reading First 3 Program
A Summary of Teacher Observation and Child Assessment Data, 2010-2013

R2LP's Early Reading First program employed a number of tools to measure teacher and student change through the course of the project. Each is described below along with highlights from a multi-year view of changes in teacher and child scores who were observed/assessed on multiple occasions between January 2010 and June 2013.

The Early Language and Literacy Classroom Observation (ELLCO) and Classroom Assessment Scoring System (CLASS) were used in teacher classroom observations by third-party evaluators from Wellesley College. The Peabody Picture Vocabulary Test, 4th Edition (PPVT-IV), Phonological Awareness Literacy Screening PreK (PALS PreK), and Test of Preschool Early Literacy (TOPEL) were the child assessment tools used with children in each of the classrooms. Student assessments were conducted by R2LP mentor-coaches and external data collectors.

TEACHER OBSERVATIONS
Early Language and Literacy Classroom Observation

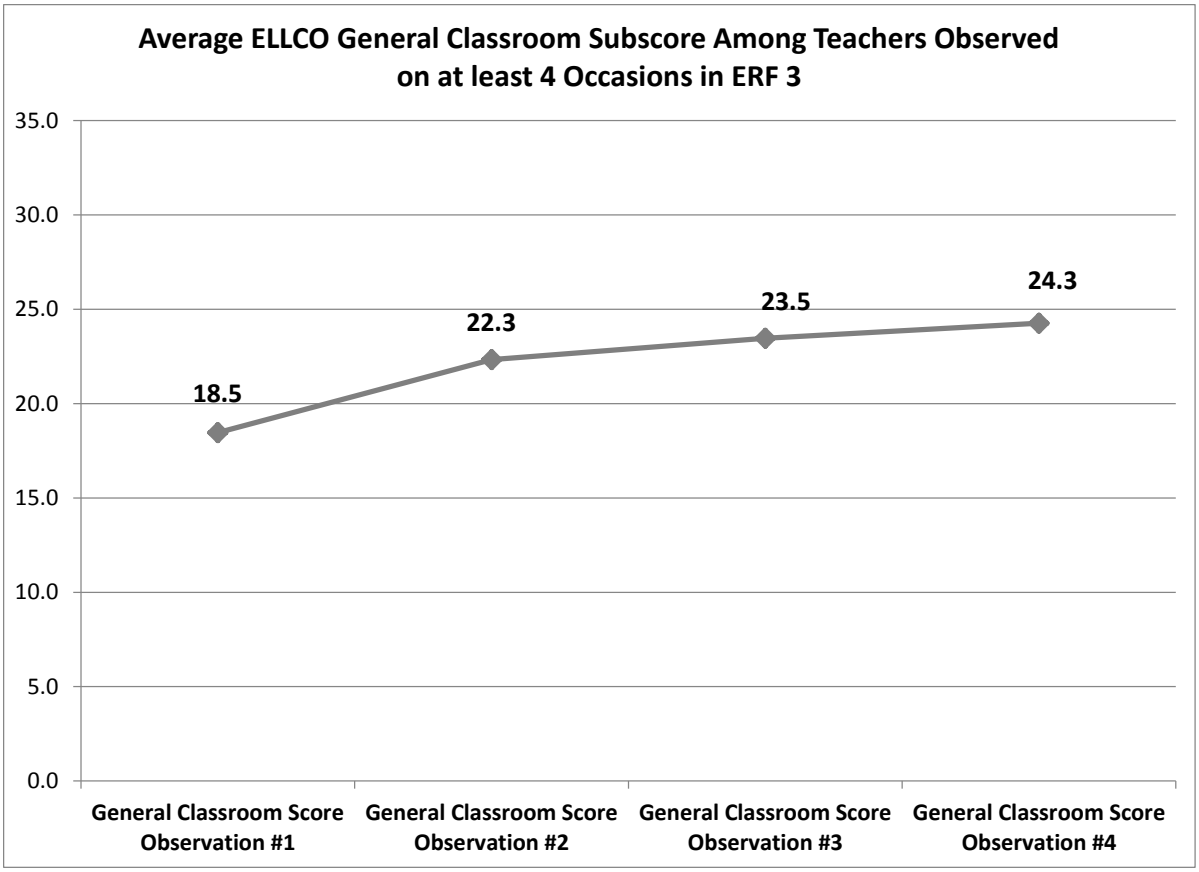
A total of 28 teachers were observed using the ELLCO instrument during ERF 3. ELLCO was used during all years of the ERF3 intervention and at each testing session occasion except for fall 2012. As shown in the following table some teachers were observed as many as seven occasions with the ELLCO, and some just a single occasion.

Number of Repeat Observations	Number of Teachers
7 Observations	4
6 Observations	2
5 Observations	3
4 Observations	6
3 Observations	1
2 Observations	2
1 Observation	10

In reviewing the collected data and assessing change over time in teachers' scores, it is a challenge to draw conclusions with a relatively small dataset, and with such small subsets of teachers observed on multiple occasions. For this reason results are included that average scores across those teachers who were observed on at least four occasions (n = 15).

The ELLCO consists of 5 sections that are summed to create two subscale scores: the General Classroom subscale and the Language and Literacy subscale. The General Classroom score is achieved by summing two observation components: Classroom Structure and Curriculum. Average scores among teachers

observed on four occasions on the General Classroom subscale are displayed below. The maximum possible score for this subscale is 35.



Teachers who were observed on at least four occasions demonstrate substantial improvement in these classroom environment domains from the first observation to those subsequent. Most notable is the sizable average increase from the first observation to the second (3.8 points). And while teachers continued to demonstrate improvements in their General Classroom scores, later changes are much smaller (1.1 points and 0.8 points, respectively, from observation two to observation three, and observation three to observation four).

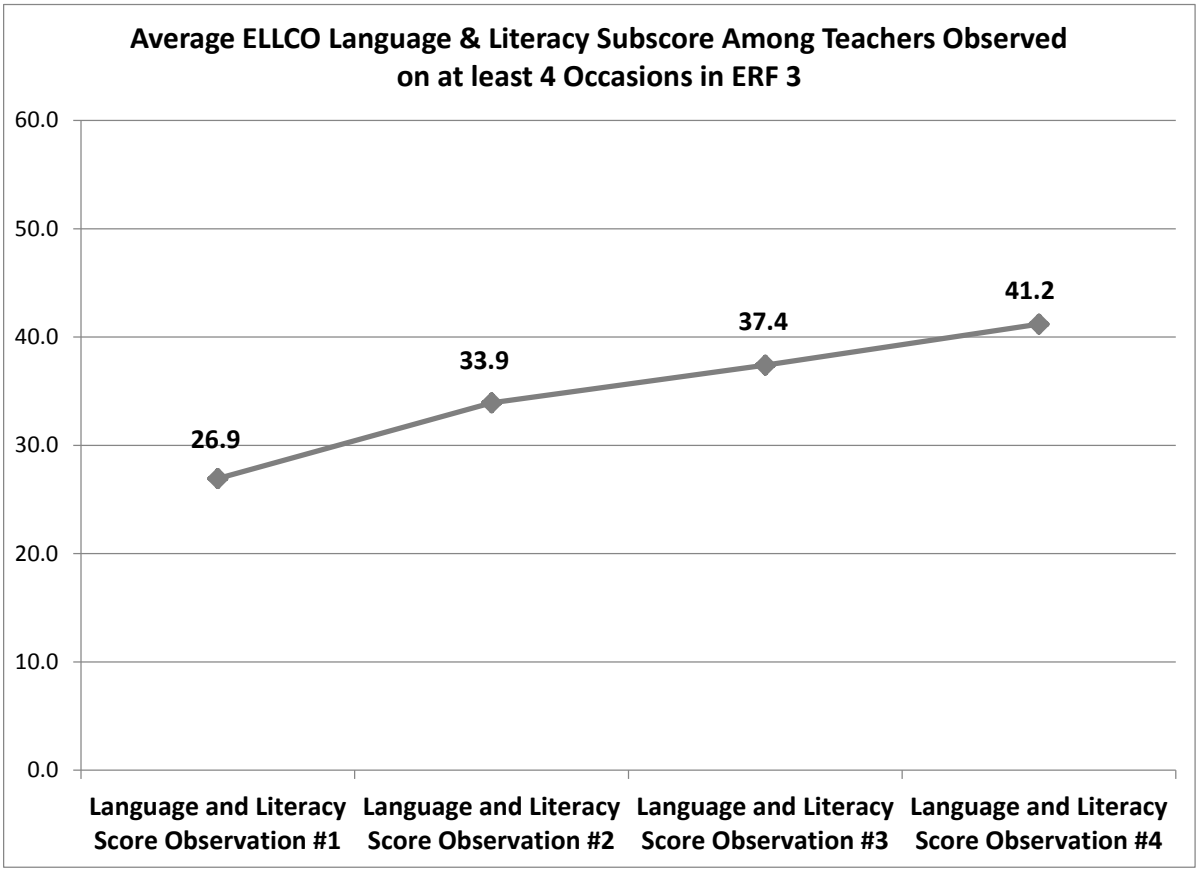
Among the smaller subset of teachers who were observed on seven occasions (*n* = 4), their average score by final observation increased to 26.0. Teachers observed on six occasions (*n* = 2) achieved a final average score of 24.5, and teachers who were observed on five occasions (*n* = 3) averaged 26.7 points.

The table below displays average scores by the number of observations for all teachers observed on three or more occasions across the ERF 3 implementation years. Interestingly, the subset of teachers who were observed on seven occasions had, on average, much lower scores at the first observation period (with a General Classroom subscore of 15.5) compared with those observed on four or five occasions (where the averages were 21.0 and 20.3, respectively). There is not uniformity in the data by center, i.e. teachers who were in the program for the longest period of time were not necessarily in the

same child care centers. Therefore averages observed are not a result of the center environment where teachers were employed.

Number of Times Observed	Average on GCE at 1st occasion	Average on GCE at 2nd occasion	Average on GCE at 3rd occasion	Average on GCE at 4th occasion	Average on GCE at 5th occasion	Average on GCE at 6th occasion	Average on GCE at 7th occasion
7 occasions (*n*=4)	15.5	19.3	22.5	25.5	25.0	23.8	26.0
6 occasions (*n*=3)	15.0	18.0	23.0	21.5	25.5	24.5	-
5 occasions (*n*=2)	21.0	26.7	27.0	29.3	26.7	-	-
4 occasions (*n*=6)	20.3	23.7	22.5	21.8	-	-	-

A similar analysis was conducted on the second ELLCO subscale, Language and Literacy, which is a score derived from summing the following sections: Language Environment, Books and Book Reading, and Print and Early Writing. Average scores across four observation periods for those teachers observed on at least four occasions are in the following figure. The maximum possible score for this subscale is 60.



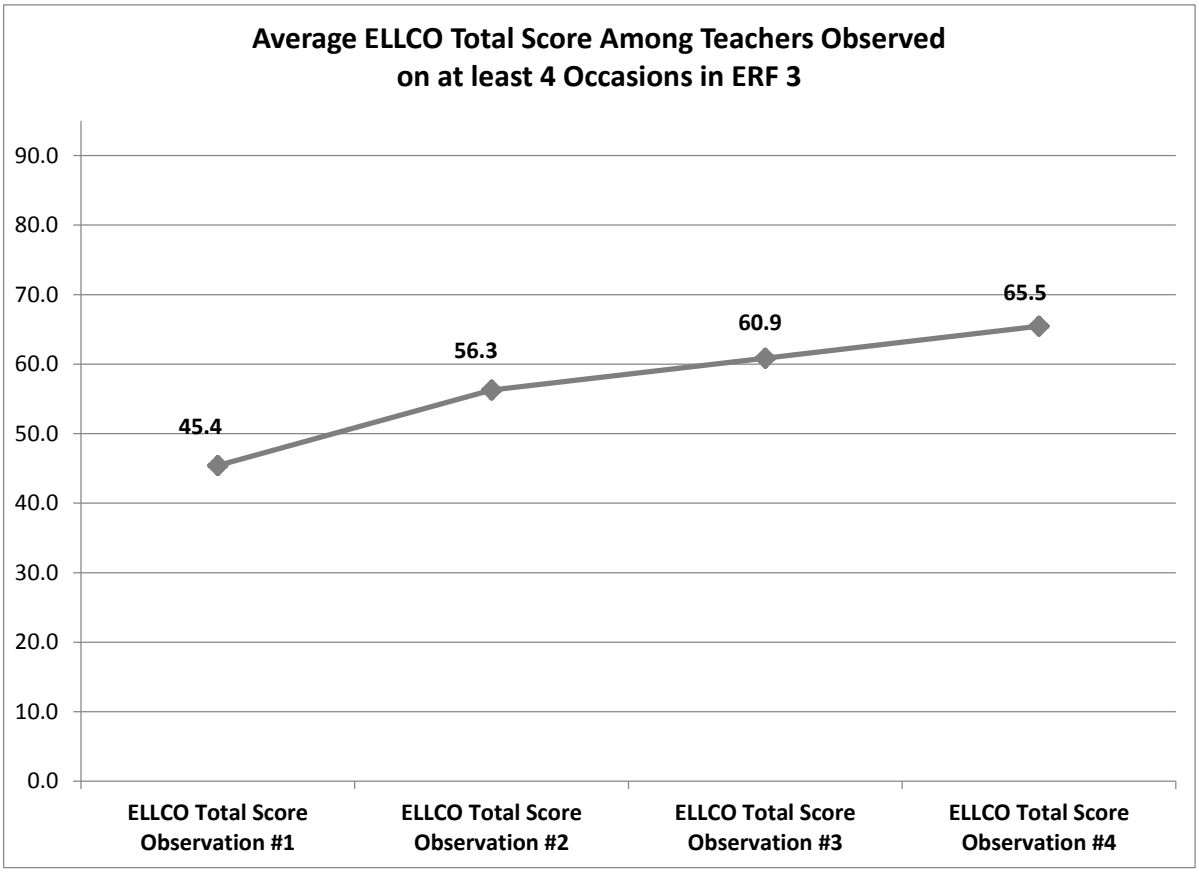
The pre-post average change on this subscale is more dramatic than that observed previously. This may be the result of the greater range of possible scores, the fact that total is derived from three observation scores as opposed to two, or a result of the content areas measured in each of these sub-scores that comprise the Language and Literacy Score. As observed with the General Classroom subscale, teachers

averaged substantial change in their scores across observation periods, with the most notable growth occurring between the first and second observations (7 point increase) compared with those in the latter observations (3.5 and 3.8 points, respectively).

The averages scores across the subset of teachers by the number of occasions on which they were observed are included in the table below.

Number of Times Observed	Average LL at 1st occasion	Average LL at 2nd occasion	Average LL at 3rd occasion	Average LL at 4th occasion	Average LL at 5th occasion	Average LL at 6th occasion	Average LL at 7th occasion
7 occasions (*n*=4)	23.5	27.3	35.5	42.3	38.8	38.3	43.5
6 occasions (*n*=3)	25.5	29.5	43.5	38.0	39.5	36.5	-
5 occasions (*n*=2)	29.5	42.0	43.0	46.3	41.7	-	-
4 occasions (*n*=6)	27.7	35.8	33.8	39.0	-	-	-

Small sample sizes complicate trending the data by the number of times that teachers were observed. For example among teachers that were observed on five occasions (n=2) the score on the fourth observation averaged 46.3, which is a much higher than the averages obtained at the third or fifth occasions for this subset of teachers. However, for purposes of demonstrating positive change from the earliest observations to those that occurred post-intervention, these averages are useful.



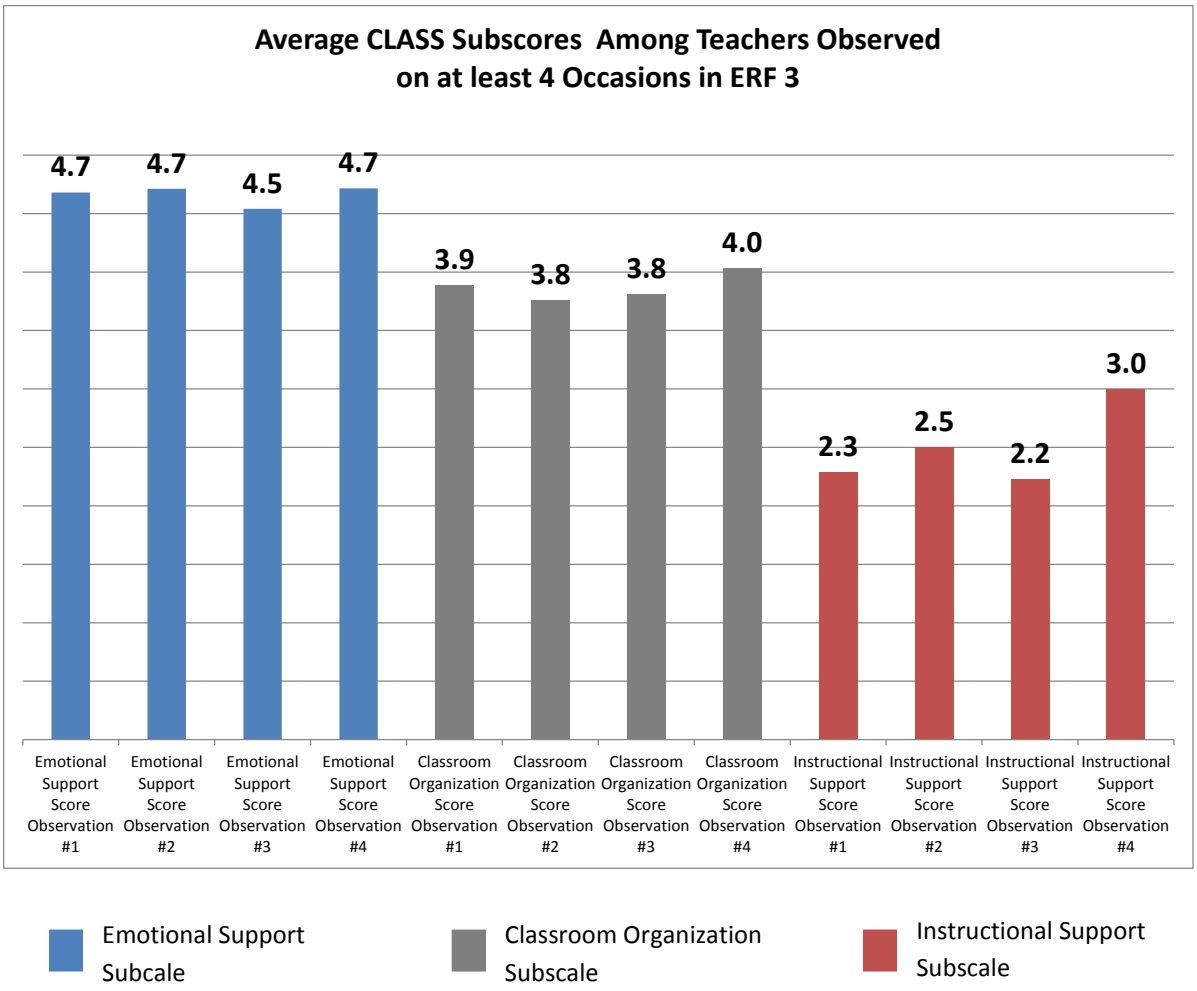
The ELLCO total score, displayed below, is the sum of the two subscores with a maximum score obtainable of 95. As demonstrated through the subscales, the overall trend in teacher scores is positive and substantial from the first to the fourth observation.

Classroom Assessment Scoring System

As with the ELLCO, R2LP used the CLASS during all years of the ERF 3 intervention and at each testing session occasion except for fall 2012. Similar with the ELLCO, therefore, some teaches have numerous CLASS tool data points across the years, depending upon how long they were involved in the ERF intervention. The maximum number of CLASS observations for any teacher is six.

Number of Repeat Observations	Number of Teachers
6 Observations	3
5 Observations	4
4 Observations	4
3 Observations	2
2 Observations	4
1 Observation	10

The CLASS consists of three subscores: Emotional Support, Classroom Organization and Instructional Support. The following charts demonstrate changes in the average scores on each for the 11 teachers who were observed with the CLASS on at least four occasions. The maximum score on each section is seven (7) points.



As displayed in the figure above, on average there was little change in the CLASS Emotional Support subscale among observed teachers from the first observation occasion to the fourth. However, these averages mask the fact that there was considerable variability in scores among teachers with a range of 3.3 to 6.6 across all results. The lowest score of 3.3 was recorded for a teacher on her first occasion being observed. The highest score of 6.6 was recorded on the fourth and final observation for a teacher. The overall average of 4.5 at the third observation is result of two particularly low scores (3.9 and 4.1) that brought the overall average down.

Average scores on testing occasion for the second subscale of the CLASS, Classroom Organization, shows little variability with overall averages for these 11 teachers always averaging between 3.8 and 4.0 points. This is out of a possible 7 points. There is sizable variation in scores, however, that is masked by displaying averages with score ranges from 2.7 to 6.0. This is true across observation periods. Even at the time of the fourth observation one teacher's classroom organization was rated 2.7 while another was rated 6.0. R2LP used the CLASS scores to tailor the interventions to meet each teacher's needs, so the variability in scores was considered and applied in the invention previously.

Finally, the Instructional Support subscale averages range from 2.3 and 2.2 on the first and third occasions, to 3.0 on the fourth occasion. Again, averages mask variation here. For example, the range of scores at the fourth observation was 1.1 (the lowest recorded score at any point among these teachers observed on four or more occasions) to 5.1, which happens to be the highest score recorded among any teacher at any time in ERF 3. Still, the average of 3.0 demonstrates considerable improvement in the Instructional Support domain among teachers.

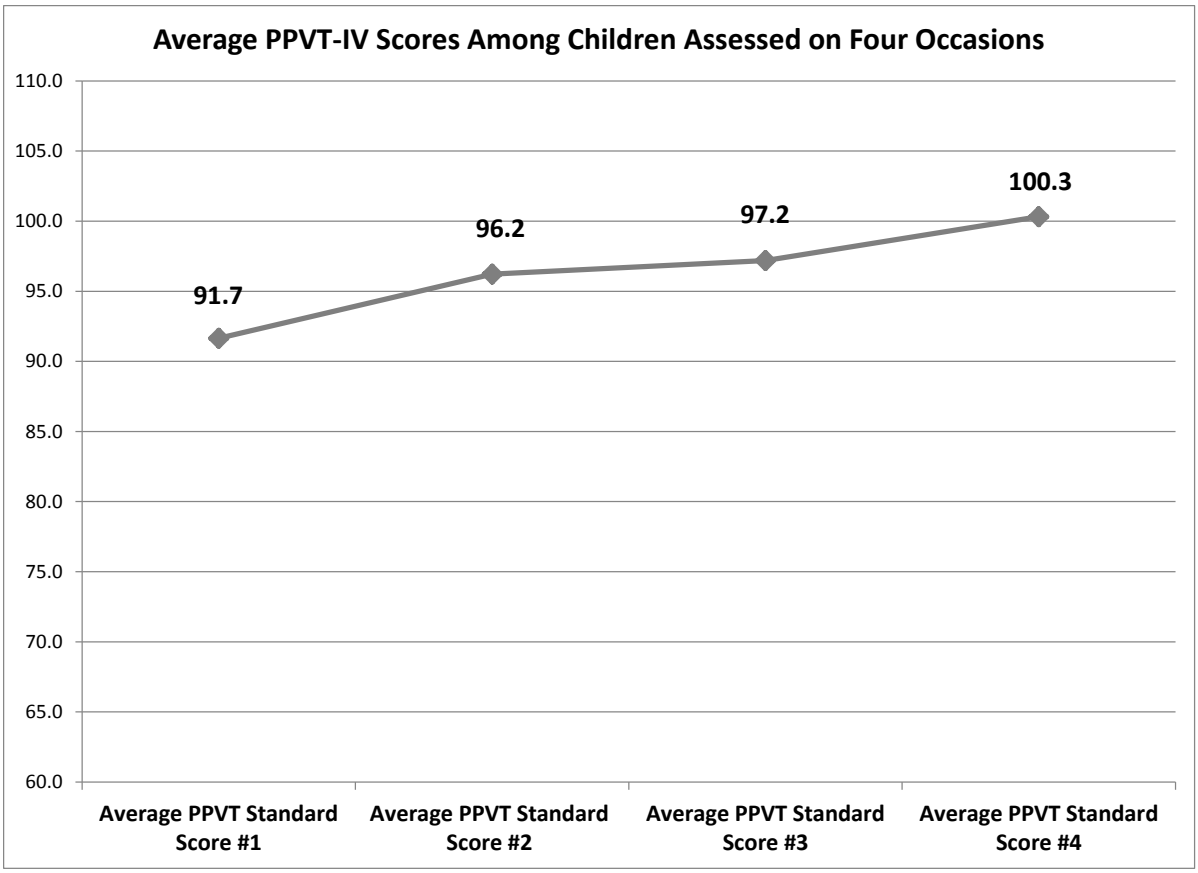
CHILD ASSESSMENTS

Peabody Picture Vocabulary Test, Version IV (PPVT-IV)

The PPVT-IV was administered to a total of 1,539 times at eight assessment occasions during the four years that R2LP operated ERF 3. A good number of children remained enrolled in ERF 3 classrooms for a full year or beyond, so the number of unique children assessed using the PPVT-IV is 666. A small subset of children remained enrolled for more than one year of the program and has four, five, or six PPVT-IV scores over time, as displayed below.

Number of Repeat Assessments	Number of Children
Children with 4 PPVT-IV scores	129
Children with 5 PPVT-IV scores	31
Children with 6 PPVT-IV scores	15

It is the subset of children with four more scores that was the focus of the longitudinal review.



The first PPVT-IV average displayed above of 91.7 is comparatively robust and well within the normal range of 85-115 for this norm-referenced assessment tool. (By comparison, the average scores among children in R2LP's ERF-II program the first time that they were assessed using the PPVT-III instrument was 86.3). As displayed above, children demonstrated sizeable increases in oral language skills through the course of ERF 3 and at subsequent times that they were assessed.

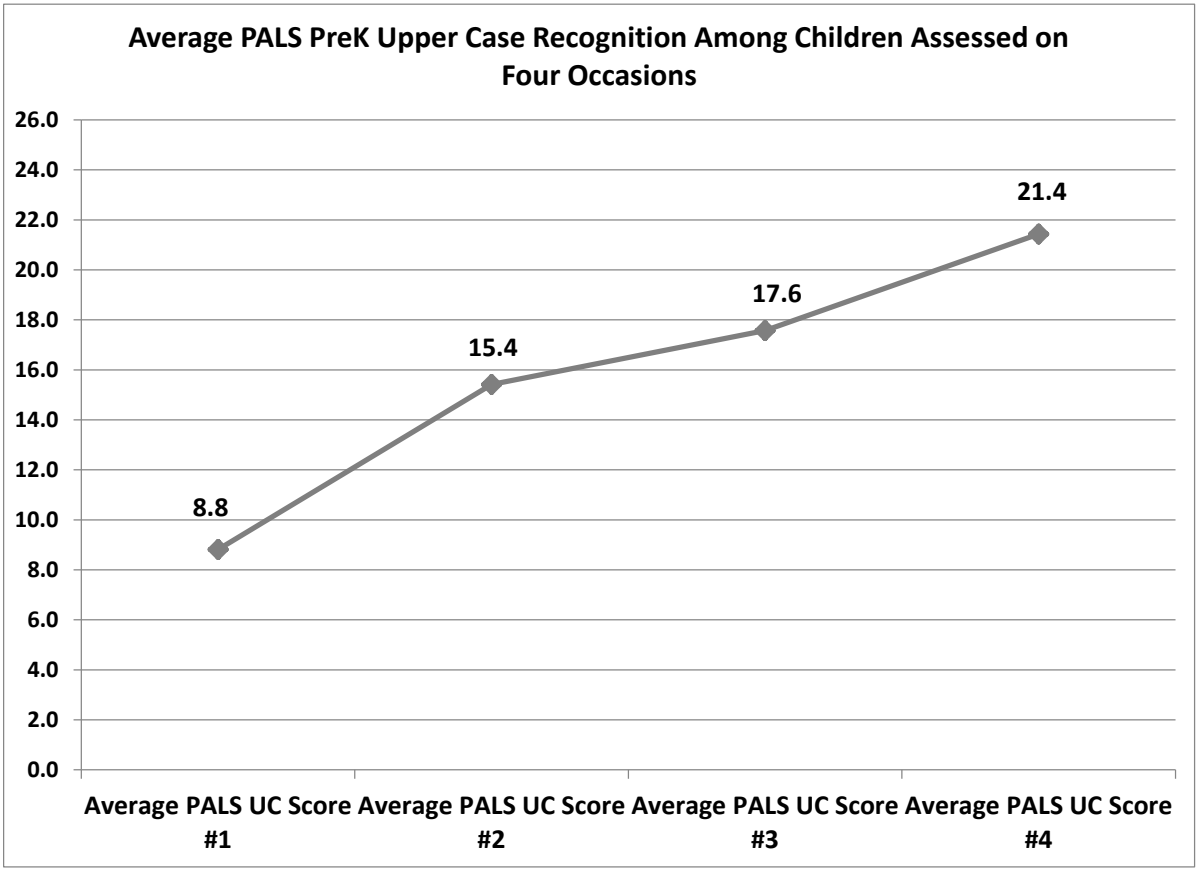
The first pre-post scores, which are typically fall to spring change, was an increase of 4.5 points on average. This reaches the significance level of growth in oral language skills in accordance with the US Department of Education's Early Reading First GPRA reporting guidelines. The second-to-third average change of just 1 point coincides with summer vacation in most cases so sizable differences wouldn't be expected (as a result of the relatively short time period of June to September, as well as parents limiting or unenrolling their children as a result of summer vacation schedules).

The second year of enrollment in the program (displayed above as "#3" and "#4") illustrates sizable change of 3.1 points on average, although below the threshold to achieve significance. Even so, an overall average of 100.3 demonstrates age-appropriate oral language skills and is more than six points above R2LP's ERF-II final average score in spring 2010 of 94.1.

For the small subset of children that were assessed on five occasions (*n* = 31), their average score on the fifth assessment occasions increases to an average of 104.4. And the 16 children assessed on six occasions averaged 105.3 on their final PPVT-IV.

Phonological Awareness Literacy Screening, PreK Edition

The PALS-PreK was administered in ERF 3 classrooms 1,541 times to 667 unique children. ERF 3 classrooms used the PALS PreK with all children and screened for literacy awareness on the following measures: Name Writing, Upper Case identification, Lower Case identification and Letter Sounds. For purposes of this longitudinal review, Upper Case Identification and a Total composite score, that is the sum of the four administered segments of the PALS PreK, are included for a subset of 134 children who were administered the PALS PreK on four or more occasions.

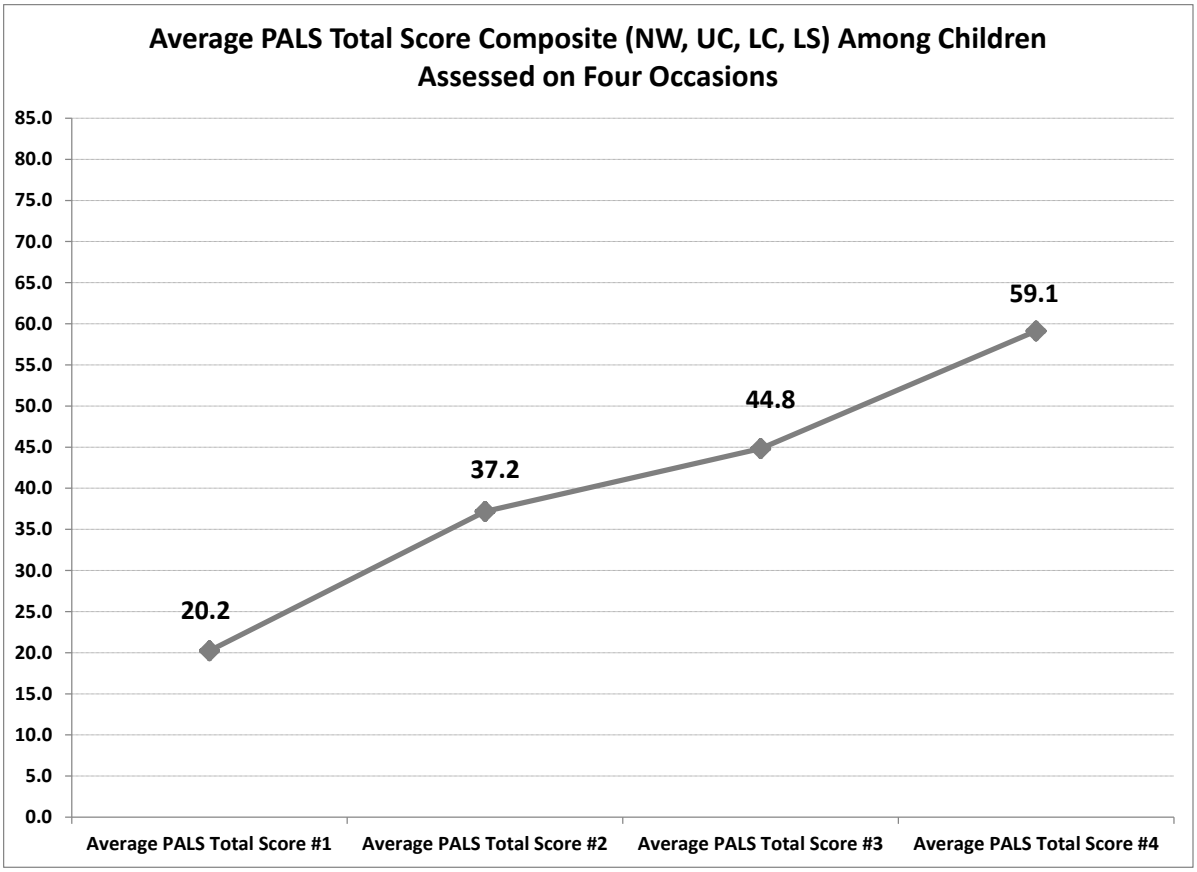


On upper case letter recognition, children made rapid progress in their first year in the program. On average, children recognized fewer than 9 letters at the outset and more than 15 at the end of the first year. Children entered their second year in the program knowing, on average, 2.2 additional letters by the time they were screened in the fall (17.6), and could recognize more than 21 letters by the end of their second year. The US Department of Education's GPRA target is 19 letters for proficiency.

A total of 28 children were screened five times with the PALS PreK and 15 children were screened six times. The average number of upper case letters identified by each group at the final assessment occasion was 22.0 and 24.4, respectively.

As noted above, R2LP administered four sections of the PALS PreK to all children, which, for purposes this analysis, were summed to reach a total score among all sections. The highest score that a child can

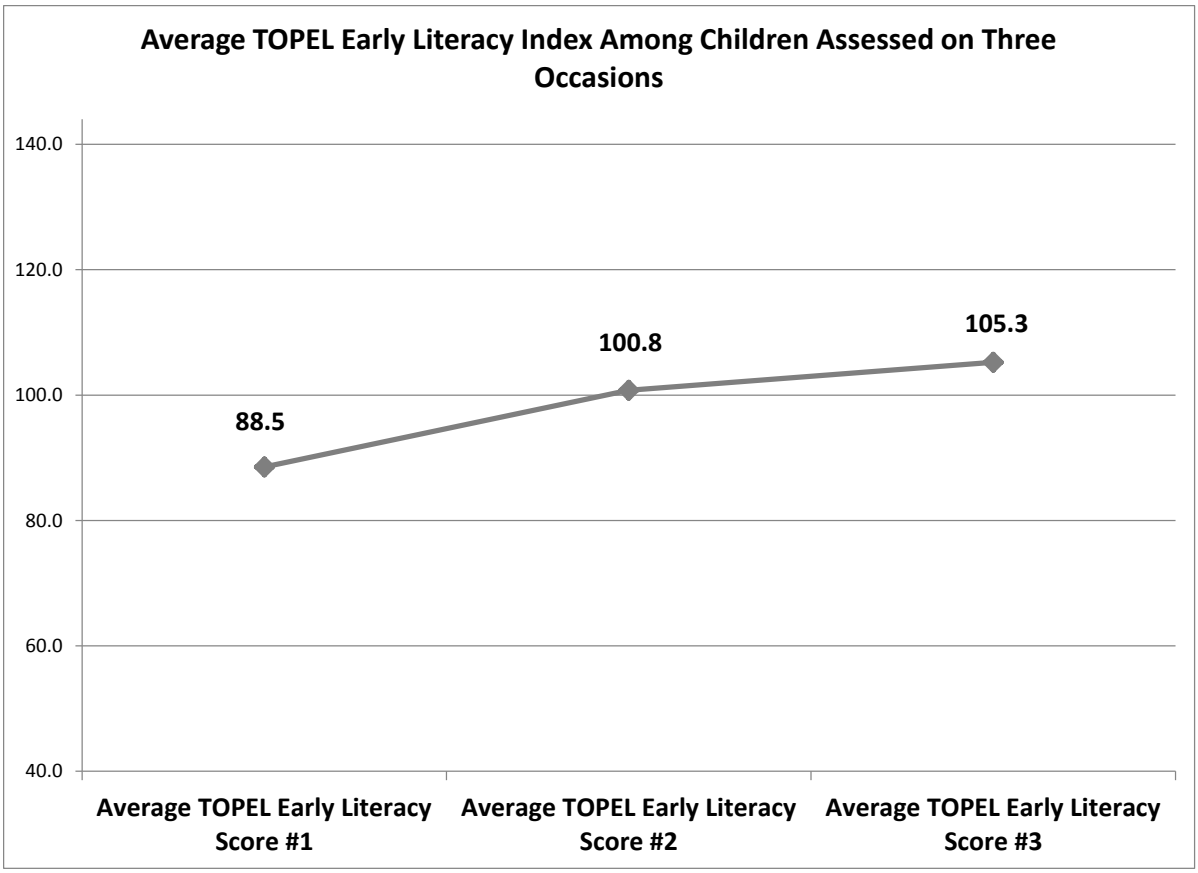
achieve is 85 across all administered section (7 points for name writing and 26 points each for upper case, lower case, and letter sounds).



In general, children demonstrated remarkable progress in their phonological awareness from pre to post each year, and across two years for this subset of children. Among the four tests that comprise this composite total, Name Writing and Upper Case identification were generally higher scores on average, while lower case and letter sounds tended to be lower. Guidance on administration of the PALS PreK encourages the administration of the lower case and letter sounds portion of the tool to children who can identify 16 or more upper case letters. Therefore R2LP's administration of all of these sections to all children regardless of their upper case knowledge would be certain to reveal children who were not developmentally able to complete these sections of the assessment tool.

Test of Preschool Early Literacy

Finally, R2LP administered the Test of Preschool Early Literacy (TOPEL) throughout the duration of the project and at all occasions except for Fall 2012. A total of 864 TOPEL results were recorded across 496 unique children. Children had, at most, five TOPEL administrations (*n*=2). Twenty-seven (27) children had four sets of TOPEL scores across multiple ERF years, and 88 have three scores. For purposes of this longitudinal review, children with at least three scores are incorporated.



At the outset, children, on average demonstrated scores within one standard deviation of the norm-referenced mean of 100, at 88.5. (R2LP's own annual benchmark for progress for annual year-end reporting purposes was achieving an average of 88.) By the time of their second assessment, children had gained an average of 12.3 points on the TOPEL and climbed to an average of 105.3 by the third assessment occasion. For the small subsets of children with four or five TOPEL scores, average changed slowed with final scores averaging 106.4 and 107.5, respectively.