Kindergarten performance in early literacy: student, teacher, and classroom factors

Michael J. Strambler[1], Joanna L. Meyer[1], Clare W. Irwin[2], and George A. Coleman[3]
[1]Yale School of Medicine
[2]Education Development Center, Inc.
[3]Cooperative Educational Services

REPORT HIGHLIGHTS

Why this study

- Resources are limited, so school districts must make decisions about a variety of teacher- and classroom-level factors that may impact the educational experience in kindergarten.
- PEER stakeholders wanted to know more about factors that might impact child performance over the kindergarten year, such as class size, teacher experience, and teacher level of education.

Study Description

- This study analyzed student data collected in Norwalk Public Schools, a mid-sized Connecticut school district, across two school years, from 2014/15 to 2015/16 (approximately 1800 students).
- PEER matched child-level data from with the Dynamic Indicators of Basic Early Literacy (DIBELS) assessment with child demographic information and teacher information.
- PEER examined the association of end-of-year DIBELS benchmark status with a variety of student-, teacher- and classroom-level factors, correcting for differences in beginning-of-year DIBELS benchmark status.

Overall trends

- Analyses revealed a low proportion of variability in DIBELS scores that could be accounted for at the classroom level.

Student-level variables

- For both study years, students qualifying for special education services were between 6 and 7 times less likely to meet DIBELS benchmark status.
- Students who were eligible for English learner services were over 2 times less likely to reach DIBELS benchmark status, but only for 2014-2015.
- Students who were eligible for free or reduced-price lunch status were 3 times less likely to reach benchmark status, but only for 2015-2016.
- Race and ethnicity were not associated with reaching DIBELS benchmark status in either year.

Teacher- and classroom-level variables

- For the 2014-2015 school year, no teacher variables predicted DIBELS benchmark status at the end of the year.
- For the 2015-2016 school year, students in classrooms with teachers holding a kindergarten endorsement were 2 times less likely to reach benchmark status. However, during this period, these teachers were also less experienced than the group of teachers without a kindergarten endorsement.
- None of the classroom-level variables were found to predict DIBELS benchmark status.

Implications

- Unusually low variation in DIBELS scores across classrooms could mean that the performance of each classroom was very similar and all teachers were administering the DIBELS accurately, or it could mean that teachers did not accurately record variation among students when administering the DIBELS.

- NPS personnel expressed some concern about the quality of DIBELS data from 2014-2015 and 2015-2016, which might explain the unusually low variation in DIBELS scores. The district has now implemented a rigorous DIBELS training program for all teachers who administer the DIBELS. It might be informative to repeat these analysis for the 2016-2017 and 2017-2018 school years.
- When considering the association of student-, teacher- and classroom-level variables with DIBELS benchmark status, only one variable demonstrated a consistent association across both study years: student eligibility for special education services.
- The relatively small number of teachers limited statistical power in the multi-level models. With a small sample, only large effects can be detected and small effects may not be measurable.
- The low variability among teachers in terms of certification status and education level may have made it difficult to assess the association of these factors with DIBELS benchmark status.
- To examine the association of student-, teacher- and classroom-level variables with kindergarten performance at a larger scale, it would be necessary to find other districts who used the DIBELS. The use of a common kindergarten assessment across districts could allow for analyses with higher statistical power.

Background

Our nation's schools operate in a world of finite resources. When making decisions about how to allocate resources, it makes sense for school districts to consider how to maximize the impact of their spending. Research has shown that factors related to a child's kindergarten experience are related to the child's long-term outcomes,[1] which suggests that kindergarten may be a particularly important year to focus on in district decision-making. Are small class sizes associated with improved child performance across the kindergarten year? When hiring kindergarten teachers, is teacher level of educational attainment associated with child performance? What is the association between teacher years of experience and child performance? Although a variety of research studies have explored such questions, it can be difficult for districts to digest this research and apply it to their own decision-making.

The Partnership for Early Education Research (PEER) was formed in 2014 to produce research that decision-makers at the local and state level can use to inform early childhood education policy and practice. Specifically, PEER aims to support decision-makers and practitioners in increasing access to high-quality early childhood education and reducing disparities in educational outcomes. (To learn more about PEER, please visit http://PEER.yale.edu.) As well as collaborating with early childhood stakeholders to develop a long-term research agenda and launch related projects, PEER has worked to explore three initial research questions. This brief focuses on the third of these initial research questions, which relates to the association of kindergarten performance with student-, teacher-, and classroom-level factors. PEER addressed this research aim by using data from one of its partner districts, Norwalk Public Schools.

Goals of the study

The purpose of the study was to understand which student-, teacher- and classroom-level factors predict kindergarten performance. Specifically, PEER explored three research questions:

1. *How do student factors such as English learner status, special education status, free or reduced-price lunch status, gender, and race/ethnicity predict kindergarten performance in literacy?*
2. *How do teacher factors such as level of education, teacher certification endorsement area, and years of teaching experience predict kindergarten performance in literacy?*
3. *How do classroom factors such as class size, percentage of students eligible for free or reduce-price lunch, and percentage of English learner students predict kindergarten performance in literacy?*

Measuring Kindergarten Performance

PEER partnered with Norwalk Public Schools (NPS) to examine these questions. Norwalk is Connecticut's sixth largest city by population and the smallest of the three PEER communities. Beginning in 2014-2015, all NPS elementary schools administer the Dynamic Indicators of Basic Early Literacy Next (DIBELS Next)[a] to students in grades K-3, a practice that grew out of Norwalk's participation in the Connecticut K-3 Literacy Initiative (CK3LI).

The DIBELS Next assessment, administered online, is a standardized direct assessment designed to measure the development of early literacy and reading skills across multiple domains. Four scales are administered in kindergarten.[2] Each scale is short, and teachers administer the DIBELS Next to one child at a time, typically on a touch-screen tablet. Administering the DIBELS Next at multiple points throughout the year provides teachers and school leaders with information about how children are progressing in the areas of language and literacy. The DIBELS Next is administered at the beginning of the year (BOY), middle of the year (MOY), and end of the year (EOY).

The four DIBELS Next kindergarten scales include:

- First Sound Fluency (FSF)—administered at beginning and middle of year (BOY and MOY);
- Letter Naming Fluency (LNF)—administered at beginning, middle, and end of year (BOY, MOY, and EOY);
- Phoneme Segmentation Fluency (PSF)—administered at middle and end of year (MOY and EOY); and,
- Nonsense Word Fluency (NWF)—administered at middle and end of year (MOY and EOY).

For each scale, the assessment developers have established a benchmark goal and cut-points for risk that can be used to assess the likelihood that an individual student will meet subsequent DIBELS Next goals. For example, if a student's score on the NWF measure is ***at or above the benchmark*** goal at the middle of the year, it is likely that with effective instruction, the student will reach the end-of-year benchmark goal. Inversely, if a student's score on the same measure is ***below benchmark*** or ***well below benchmark*** at the middle of the year, it is unlikely that the student will reach the end-of-year benchmark goal without strategic or intensive intervention.

For each timepoint, a student's scores on the DIBELS Next scales can be combined to calculate the DIBELS Composite Score, which serves as an overall estimate of the student's early literacy and reading skills. DIBELS Composite scores can also be compared to relevant benchmark goals and cut-points for risk, allowing teachers to identify which children are at or above benchmark, below benchmark, or well below benchmark and individualize instruction accordingly.

The four DIBELS Next scales have strong evidence of reliability; interrater reliability for all measures is above .94, with the internal consistency for the composite score equal to .83. Concurrent and predictive validity was assessed against the Group Reading Assessment and Diagnostic Evaluation (GRADE). Concurrent validity estimates range from .24 for Phoneme Segmentation Fluency to .40 for NWF Correct Letter Sounds, and the composite score has a concurrent validity measure of .40. Predictive validity

[a] Dynamic Measurement Group, the authors of the DIBELS Next® assessment, announced in October 2018 that the DIBELS Next has been renamed Acadience™ Reading. For more information see https://acadiencelearning.org/ann_acadience.html.

estimates range from .19 for NWF Whole Words Read to .52 for First Sound Fluency, and the composite score has a predictive validity measure of .48.[3]

How the study was conducted

For this study, Norwalk Public Schools, Yale University and EDC executed a written data sharing agreement, which defined the goals of the study, what data would be needed, and how student privacy would be protected. Once this document was in place, Norwalk Public Schools shared student data and teacher data for all 2014/15 and 2015/16 kindergarten students and kindergarten teachers. The student data included English learner (EL) status, special education status, free or reduced-price lunch (FPRL) status, gender, race/ethnicity, teacher name, and DIBELS data from each of three assessment timepoints. Specifically, DIBELS data included numeric scores (continuous variables) and benchmark scores (categorical variables) for all kindergarten students assessed at the beginning, middle or end of the year. PEER used student data to calculate values for three classroom-level variables: class size, percentage of FRPL-eligible students, and percentage of EL students. The teacher data included level of education (for example, master's degree, sixth year certificate, etc.), teacher certification endorsement area (for example, Elementary K-6, Bilingual PK-12, etc.), and years of teaching experience in the district.

After linking the student and teacher data, PEER researchers conducted analyses using multilevel modeling (MLM), which is designed for handling nested data such as students grouped within classrooms and schools. In other words, an MLM approach is more appropriate than traditional regression analyses because students' scores on a given measure may be partially dependent upon their teacher, classroom grouping, or school.

For each MLM model, PEER researchers used the teacher ID as a grouping variable and end-of-year (EOY) DIBELS benchmark status as a binary outcome (whether or not a student met the benchmark). Student demographics and beginning-of-year (BOY) DIBELS benchmark status were entered as level 1 variables (student-level indicators.) Specifically, level 1 variables included EL status, special education status, FPRL status, gender, race/ethnicity, and T1 DIBELS benchmark status. Teacher-level variables were entered as level-2 variables in the models. These variables consist of a continuous measure of teachers' number of years of experience, an ordinal measure of teachers' level of education, and a binary variable describing whether a teacher had at least one teacher certification that included kindergarten.

PEER researchers conducted several different MLM models based on this structure, each of which was conducted separately for the 2014-2015 school year and the 2015-2016 school year. We first conducted an analysis for each year to determine the intraclass correlation (ICC), which describes the proportion of variability in DIBELS scores that can be accounted for at the classroom level. Second, we conducted a model with just the student-level demographic variables (level 1) and examined which of these factors were associated with EOY DIBELS benchmark status, controlling for BOY DIBELS benchmark status.

The third and fourth models included the teacher-level variables (level 2) along with the student-level demographic variables (level 1), again controlling for BOY DIBELS benchmark status. Finally, we examined whether classroom-level factors contributed to students' EOY DIBELS benchmark status. For this analysis, we entered classroom-level variables (class size, percentage of FRPL-eligible students, and percentage of EL students) as level-2 variables and students' BOY

DIBELS benchmark status as a level-1 variable in a random-intercepts model. See Appendix A for more details regarding the study methodology.

Sample

The total sample population for this study consists of 1,792 kindergarten students who were enrolled in Norwalk Public Schools at the close of the 2014/15 or 2015/16 school year and for whom the district provided DIBELS results or demographic information. Students repeating kindergarten during either year were included in the sample, because PEER could not identify repeaters for the 2014/15 school year without 2013-2014 data. The number of repeaters in 2015/2016 was small (1.6 percent).

Two hundred and sixty-five (265) kindergarten students were excluded from analysis because they were missing beginning of year (BOY) and/or end of year (EOY) DIBELS data or because their data couldn't be linked to teacher data. This resulted in a sample size of 1,527 of which 751 were from the 2014-2015 school year and 776 were from the 2015-2016 school year. The excluded students may be systematically different from the overall population for a variety of reasons; for example, it is possible that these students were not present in the district for the entire year. For more details on the sample, see Tables B.1a-B.1e in Appendix B.

According to the demographic and DIBELS data, the kindergarten students in the sample were taught by 53 teachers, 44 of whom taught kindergarten in 2014-2015 school year and 45 of whom taught kindergarten in 2015-2016. Degree status, certification endorsement area, and years of teaching experience were available for 51 of these 53 teachers. The remaining two teachers included one special education teacher whose self-contained classroom included very few kindergarteners and one long-term substitute teacher. See Tables B.2a through B.4b in Appendix B for more information on the teachers included in the study.

What the study found

Overall Trends

Of the kindergarten students included in analyses for the 2014-2015 school year, 58.6% had composite scores that were considered at or above benchmark at the start of the school year, compared to 82.6% at the end of the school year. For the 2015-2016 school year, 62.1% met benchmark at the beginning of the year, whereas 80.0% met benchmark by the end of the school year. The mean values for composite scores were 34.04 and 148.54 at the beginning and end of 2014-2015, respectively, compared to 35.66 and 150.01 at the beginning and end of 2015-2016. For more information about DIBELS data, see Tables C.1 and C.2 of Appendix C.

For each year, the unconditional multi-level model produced a very low value for the interclass correlation (ICC), which describes the proportion of variability in DIBELS scores can be accounted for at the classroom level. The ICC of 8% for 2014-2015 and 9% for 2015-2016 are quite small relative to what other research involving the DIBELS has found. For example, one study found an average ICC of 35%.[4] In the Discussion section, we discuss some factors that may contribute to this low variability.

Of the 51 teachers for whom data were available, all had attained a master's degree or higher. Of the 44 kindergarten teachers in 2014-2015 for which data was available, 84.1% held a teacher certification endorsement that included kindergarten and 15.9% did not, compared to 86.7% and 13.3%, respectively, for 2015-2016. When considering the same teachers, the mean teaching experience was 12.9 years in 2014-2015, compared to a mean of 12.2 years for 2015-2016. See Table C.3 of Appendix C for more information.

Student-level factors and performance

For the 2014-2015 school year, qualifying for special education services was associated with lower odds of reaching DIBELS benchmark status (OR=.181, b =-1.711, $p < .001$). Students who were eligible for special education services were

nearly 6 times less likely to reach benchmark than students who were not eligible for special education services. Qualifying for English learner services also was associated with lower odds of reaching DIBELS benchmark status (OR=.480, *b* =-0.735, *p* = .021) for 2014-2015. English learner students were over 2 times less likely to reach benchmark status than non-English learners. There was no association between free or reduced-price lunch status and DIBELS benchmark status in 2014-2015. See Table C.4a in Appendix C for the results of this analysis.

For the 2015-2016 school year, qualifying for special education services again was associated with lower odds of meeting DIBELS benchmark (OR=.146, *b* = -1.925, p <.001). Similar to 2014-2015, students eligible for special education services were nearly 7 times less likely to reach DIBELS benchmark status than those not eligible for special education services. Qualifying for free lunch status was associated with lower odds of reaching benchmark status (OR=.344, *b*=-1.066, *p* < .001) for 2015-2016, with free lunch-eligible students being nearly 3 times less likely to reach DIBELS benchmark status than those who were ineligible for free lunch. There was no association between English learner status and DIBELS benchmark status in 2015-2016. See Table C.4b in Appendix C for the results of this analysis.

As shown in Tables C.4a and C.4b, there was no association between race/ethnicity and DIBELS benchmark status in either year, nor was there any association between gender and DIBELS benchmark status in either year.

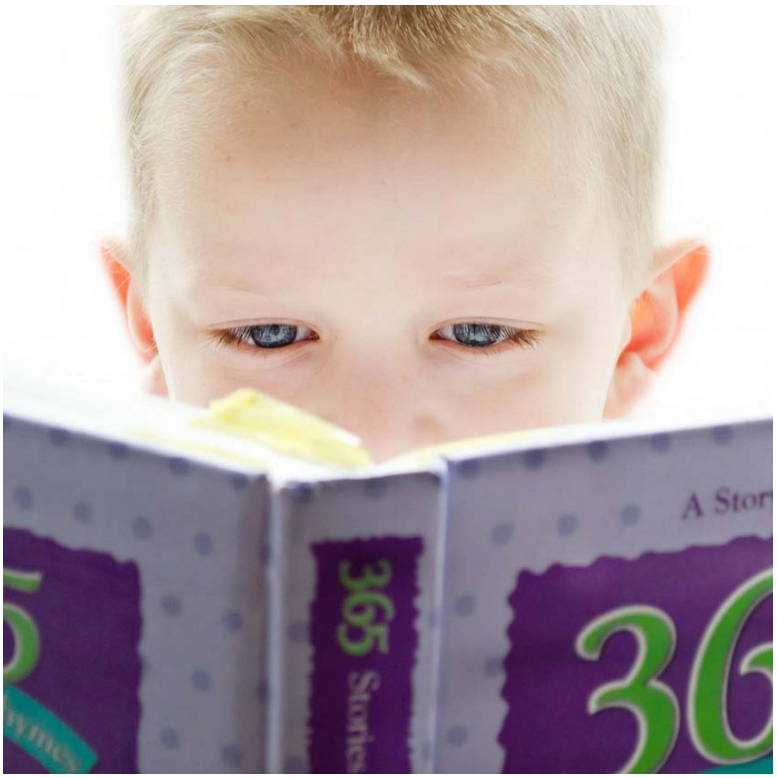
Teacher-level factors and performance

For the 2014-2015 school year, no teacher variables predicted DIBELS benchmark status at the end of the year. (See Table C.5a.) In addition, no teacher variables were associated with the link between special education and DIBELS benchmark status, and no teacher variables were associated with the link between English language status and DIBELS benchmark status. (See Table C.6a.)

For the 2015-2016 school year, certification endorsement area was associated with lower odds of reaching benchmark status (OR=.391, *b*=-.940, *p*=.024). Students in classrooms with kindergarten-endorsed teachers were over two times less likely to reach benchmark status than those is classrooms with teachers without a kindergarten endorsement. (See Table C.5b.) We observed that in 2015-2016, the kindergarten-endorsed teachers were notably less experienced than the non-kindergarten endorsed teachers (see Table C.9a), as described when we further discuss this counterintuitive result below. As in 2014-2015, no teacher variables were associated with the link between special education and DIBELS benchmark status or the link between free lunch status and DIBELS benchmark status. (See Table C.6b.)

Classroom-level factors and performance

The models for this analysis included three classroom-level factors: class size, percentage of students qualifying for free or reduced priced lunch, and percentage of students qualifying for EL services. When we included individual-level student demographics in the models as control variables (specifically race/ethnicity, EL status, and FPRL status), none of the classroom-level variables were found to predict DIBELS benchmark status. For results of these analyses, see Tables C.7a, C.7b, C.8a, and C.8b in Appendix C.)



Discussion

This study examined the association of teacher characteristics, student characteristics, and classroom factors with kindergarten performance as measured by the DIBELS assessment. Two consistent findings emerged. First, DIBELS data for this sample demonstrate unusually low variation among classrooms for both 2014-2015 and 2015-2016, raising questions about the quality of the data during the study years. Second, special education status was associated with lower odds of reaching DIBELS benchmark status in both years, such that students in special education were between 6 and 7 times less likely to reach DIBELS benchmark status than their general education counterparts.

Some results were found inconsistently across the two years. For example, English Learner status was found to be associated with DIBELS benchmark status in one year but not the other, and the same is true for free or reduced-price lunch status. Given this inconsistency, we advise caution in drawing firm conclusions from these results. The consistent results noted above are more likely worthy of attention.

In terms of the surprisingly low variation in DIBELS scores among classrooms, teachers may have provided such similar ratings across classrooms because they were very well trained on the administration of the DIBELS measures. Conversely, it is possible that teachers did not accurately record variation among students when administering the DIBELS, but instead scored them according to some other common expectation, such as how they "should" score students at any given timepoint.

In order to gain more context for this finding, PEER consulted with the district's instructional specialist for K-12 English language arts to learn more about DIBELS administration. We learned that when the instructional specialist entered her position in 2015-2016, she noted that DIBELS scores appeared to be inflated. Although the majority of students received ***at or above the benchmark*** ratings, a significant number of those students were not performing at grade level in the classroom. This information, along with the above finding, led us to surmise that there may be some issues with data quality resulting from teachers not accurately capturing student variation during the study years.

It is encouraging that the district responded to concerns about the accuracy of DIBELS data by implementing a rigorous training program that includes a full-day training for all new teachers and calibration booster sessions for all teachers ahead of each DIBELS administration. In these booster sessions, teachers work in small groups to score simulations until their scores are within two points of an expert score.

Regarding teacher-level predictors of DIBELS benchmark status, we found that neither the number of years teaching in the district nor teacher level of education were predictive of DIBELS benchmark status. This finding may be due to the restricted range of data on these teacher characteristics. Namely, all of the teachers in the sample had a master's degree or higher and there was also limited variation in teachers' years of experience.

One surprising finding was that in 2015-2016, teacher certification endorsement for kindergarten was associated with lower odds of reaching DIBELS benchmark status. We viewed this as a counterintuitive finding since we would expect that students in classrooms with kindergarten-endorsed teachers would be more likely to reach benchmark status than those in classrooms with teachers who are not endorsed to teach kindergarten. However, it is important to note that in

2015-2016, teachers with a kindergarten endorsement were less experienced as a group than teachers without a kindergarten endorsement, whereas both groups of teachers had similar years of experience in 2014-2015, as shown in Tables C.9a and C.9b. As such, the association of kindergarten endorsement with DIBELS benchmark status may actually be a function of the experience of those teachers who are permitted to teach kindergarten without holding a kindergarten endorsement.

It is also important to note that the relatively small number of teachers may have limited our ability to detect effects at the teacher level. As sample size decreases, the ability to detect effects also decreases. In other words, with a small sample, only large relationships between factors can be detected and small effects may be missed. For instance, teacher years of experience may have an association with child outcomes that is too small to detect with a classroom-level sample of this size.

Implications and Next Steps

The findings have several implications for further research on this topic and for applied education research in general. First, as described in the October 2017 PEER brief on the use of early childhood assessments, it is important to consider the match between the desired use and the intended use of an assessment, as defined by its developers. Although studies have documented the reliability and validity of the DIBELS, as well as their sensitivity to student change, the DIBELS is not designed to measure individual students' growth across the year. This fact means that the DIBELS may not be an effective measure for assessing the relationship of instructional, curricular, classroom, or teacher factors with student achievement. As districts identify their assessment goals and select assessment tools, researcher partners may be able to offer strategic support.

This study raised some questions around data quality, which point to the importance of training teachers on the systematic, objective use of assessments. To prepare teachers to administer the DIBELS accurately, Norwalk Public Schools decided to increase the rigor of DIBELS training in 2016-2017, a change that appears to be a strong and warranted step toward data quality improvement. The district has demonstrated its commitment to data quality by providing training for all new teachers and conducting booster sessions before each DIBELS administration. These training opportunities are especially important given the mobility of teachers among schools and across grade levels. An interesting follow-up study might be to analyze the data from the 2016-2017 and 2017-2018 school years and evaluate whether the findings were substantially different from those indicated in this report, which used 2014-2015 and 2015-2016 data.

This study was limited by the absence of a common early childhood assessment that can be used to examine student growth across communities. A statewide early childhood assessment could increase opportunities to generate evidence on the effectiveness of educational practices, especially if assessment data were stored electronically in a central database. Such evidence could inform decision-making around instructional and programmatic approaches, as well as resource allocation.

Terminology

Intraclass correlation coefficient (ICC), in the context of this report, describes how much variability in DIBELS scores is related to the grouping of students in classrooms. This is expressed as a proportion of the variability at the classroom level divided by the variability at the classroom level plus the variability at the individual level.

Multilevel modeling (MLM) is a statistical approach that accounts for interdependence, or clustering, among data. Traditional regression analyses assume that each student's score on a measure is independent of other students' scores. When data demonstrates shared variation as a result of the grouping of students within classrooms and/or schools, multilevel modeling is a more appropriate approach than traditional regression analysis.

Acknowledgements

This work was supported by grant R305H140142 from the U.S. Department of Education's Institute of Education Sciences (IES) under the Researcher-Practitioner Partnerships in Education Research program. The content of this brief is solely the responsibility of the authors and does not necessarily represent the official views of the Institute of Education Sciences (IES), the U.S. Department of Education, the Connecticut State Department of Education (CSDE), the Connecticut Office of Early Childhood (OEC), or Norwalk Public Schools (NPS).

PEER would like to thank the NPS representatives who made this study possible, including Ralph Valenzisi, Cathy Orgovan, Jean Starkman, Jamie Zyla, Pamela Augustine-Jefferson, Janine Goss, Dr. Brenda Myers, and Dr. Steven Adamowski. PEER would also like to thank Dr. Nicholas Gage from the University of Florida for his contributions to the project, as well Mary Oster and Diane Beltz-Jacobson from the City of Norwalk for their support. Finally, PEER would like to thank PEER interns Silvana Alarcon and Sabrina Trombetta.

References

1. Chetty, R. et al. (2011) How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, **126**(4): p. 1593–1660. https://doi.org/10.1093/qje/qjr041
2. Dynamic Measurement Group. (2018) DIBELS Next Information Sheet. https://acadiencelearning.org/DIBELS_Next_Info.pdf
3. Dewey, E. N., Powell-Smith, K. A., Good, R. H., Kaminski, R. A. (2015) *DIBELS Next Technical Adequacy Brief*. Eugene, OR: Dynamic Measurement Group, Inc. https://acadiencelearning.org/papers/Acadience_Reading_Technical_Adequacy_Brief.pdf
4. Waterman, C., et al. (2012) The matter of assessor variance in early childhood education—Or whose score is it anyway? *Early Childhood Research Quarterly*, 27(1): p. 46-54. https://doi.org/10.1016/j.ecresq.2011.06.003

Appendices

All appendices for this report are available on Open Science Network by visiting https://osf.io/urwmf/.