HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2019 No. 003

An Impact Evaluation of *Supplemental Blended* Implementation for Mathematics at Grades 6–8

Final Report

Prepared for:
Curriculum Associates
153 Rangeway Road
North Billerica, MA 01862

Authors:
Rebecca Norman Dvorak
Bruce Randel (Century Analytics)

Date:
April 9, 2019

An Impact Evaluation of *Supplemental Blended* Implementation for Math at Grades 6–8

Table of Contents

List of Tables

An Impact Evaluation of Supplemental Blended Implementation for Mathematics at Grades 6–8

Introduction

Founded in 1969, Curriculum Associates provides a variety of educational products and services with the goal of improving education for students and teachers. Three Curriculum Associates products available for mathematics and reading include *i-Ready® Diagnostic* (available for K–12), *i-Ready® Instruction* (available for K–8), and *Ready® Supplemental Curriculum* (available for K–8). The *i-Ready Diagnostic* (a) are online, computer adaptive assessment that pinpoint student needs at the sub-skill level in reading and mathematics and (b) help monitor the extent to which students are on track to achieve end-of-year targets. The *i-Ready Instruction* program provides online, individualized instruction for students and is designed for use with *i-Ready Diagnostic*. Curriculum Associates *Ready* curriculum materials are accessible through an online *Teacher Toolbox* system or through print materials. The intent of this curriculum is for a school or district to use these materials in conjunction with a separate, primary curriculum. When *i-Ready Diagnostic*, *i-Ready Instruction*, and the supplemental use of *Ready* are used together they are referred to as a *Supplemental Blended* implementation.

Curriculum Associates' *i-Ready Diagnostic* is currently used by 5.6 million K–12 students across the United States. Some of these students use the Diagnostic only, while others also use the Online Instruction and/or other components designed to enhance student learning. Though not as widespread, the *Ready* curriculum is used for supplemental instruction across the United States in conjunction with core curriculums to provide additional instruction by educators, as they see fit.

The Human Resources Research Organization (HumRRO) conducted an evaluation to examine the impact of the *Supplemental Blended* implementation on student achievement in mathematics for grades 6–8 compared to use of *i-Ready Diagnostic* only. This study was designed to meet the required rigor of the What Works Clearinghouse (WWC) 4.0 standards to achieve a rating of *Meets WWC Group Design Standards with Reservations (*WWC, 2017), and to meet guidelines for a Level 2 (or *Moderate*) rating for the Every Student Succeeds Act (ESSA) guidance for evidence-based research (U.S. Department of Education, 2016). This was achieved by using a quasi-experimental design (QED) which involved establishing baseline equivalence between a treatment and comparison group, using an outcome measure acceptable by WWC, including baseline achievement as a covariate, and sampling in a manner that avoids confounding variables.

Research Questions and Study Design

This evaluation addressed the following research question for students in grades 6–8.

1. Does use of *Supplemental Blended* implementation for mathematics result in higher student achievement on *i-Ready Diagnostic* beyond use of *i-Ready Diagnostic* only?

In this study, our hypothesis was that student achievement would benefit at schools using a *Supplemental Blended* implementation for mathematics over use of *i-Ready Diagnostic* only. This hypothesis is based on the belief that students benefit from the *i-Ready Instruction* targeted to their specific needs, and from incorporation of *Ready Mathematics* in classroom instruction. We predicted that the use of these products benefits student achievement in mathematics and will be reflected in *i-Ready Diagnostic* scores in the spring following one school-year of use.

Cluster-Level Design

Assignment into the treatment or comparison group occurred at the school level. Although a district may choose to implement one or more Curriculum Associates products district-wide, data and discussions with those who work directly with districts indicated there is great variation between schools regarding whether and how implementation occurs. Thus, we determined the appropriate unit of assignment for this study was the school. The unit of observation in this study was the student, with student-level achievement on the *i-Ready Diagnostic* serving as a baseline matching measure and outcome measure.

Outcome Measure

Both the treatment and comparison group administered the *i-Ready Diagnostic* for mathematics in fall and spring during the study period. The *i-Ready Diagnostic* was designed to be aligned to common state standards and was administered to students across the country. Because the *i-Ready Diagnostic* for mathematics measures achievement aligned to common mathematics content and skills with proven reliability (see Appendix A), the assessment met the guidance set forth by the WWC 4.0 standards for an acceptable outcome measure (WWC, 2017). There was the added benefit that the measure is given early and late in the school year, allowing a consistent baseline and outcome achievement measure.

Identifying a Sample of Schools

HumRRO worked with Curriculum Associates to identify a sample of schools with students in grades 6–8 that implemented *Supplemental Blended* for mathematics. We first identified potential districts through Curriculum Associates' *i-Ready* usage data and *Ready* purchase data - the *Ready* purchase data specified those districts and schools that had purchased *Ready* print books and/or *Teacher Toolbox*. We then examined *i-Ready Diagnostic* and *Instruction* use for the schools within these districts to determine which schools meet our definition of full *Supplemental Blended* implementation. Through our review of *Ready* purchase data and *i-Ready* student usage data, we determined most *Supplemental Blended* schools used the program for both reading and mathematics. Therefore, we decided to include only those schools with reading and mathematics usage.

For comparison schools, we identified districts including schools with students that had used *i-Ready Diagnostic* only, with no *i-Ready Instruction* use. We then checked that these districts (and the specific schools within districts) had not made *Ready* print or *Teacher Toolbox* purchases. We decided to include this same restriction for comparison schools as we did with treatment schools by selecting only schools with students that had used *i-Ready Diagnostic* for reading and mathematics. This allowed reading and mathematics achievement to be used for student-level matching to gain a more similar set of treatment and comparison students than what would have been possible with mathematics scores alone.

The following summarizes the qualifications for school-level participation in this study for our treatment and comparison groups:

- ***Supplemental Blended* (treatment):** School that used *i-Ready Diagnostic*, *i-Ready Instruction*, and *Ready* curriculum as a supplemental program for reading and mathematics. *Ready* use included those that used print materials and those that used the online *Teacher Toolbox* as supplemental to a separate core curriculum. Schools selected as eligible for this group had students that used *i-Ready Diagnostic* and *i-*

Ready Instruction across grades. Only those using all three products for reading and mathematics were included.

- ***i-Ready Diagnostic* only (comparison):** Schools that administered the *i-Ready Diagnostic* at minimum of fall and spring for reading and mathematics. These schools did not use *i-Ready Instruction* or *Ready* as a supplemental or core curriculum for reading or mathematics. Schools selected as eligible for this group came from districts that did not make any *Ready* purchases at or before the time studied.

Through discussions with Curriculum Associates and review of data, we determined the treatment start period, defined as the point of implementing *Supplemental Blended*, was variable, with eligible schools beginning treatment either in the 2014–15, 2015–16, or 2016–17 school year. Similarly, the comparison group included schools that started *i-Ready Diagnostic* use in various years. Though we included only one year of data for each school, we incorporated a cohort variable to ensure treatment schools were matched to schools within their cohort. The cohort variable was also used in our impact analysis as a covariate.

Analysis

Achieving Baseline Equivalence

Once schools meeting the requirement for each group were determined, matching was conducted to identify the comparison group of students from the identified schools. Matching was conducted at the student level on student achievement using fall *i-Ready Diagnostic* reading and mathematics scores. Though the study outcome measure is mathematics only, including reading *i-Ready Diagnostic* scores as a second baseline measure provided additional information to ensure students were matched with a similar-ability student.

Logistic regression was used to compute a propensity score for each student in the treatment and the comparison groups. The model predicted the chance that each student belonged to the treatment group through a propensity score between 0 and 1. Student variables used to determine these scores were the mathematics and reading *i-Ready Diagnostic* scores for the fall of the first year of treatment implementation (fall 2014, 2015, or 2016, depending on the cohort). Matching was stratified by grade level and cohort to ensure all students were matched with someone at the same grade level and for the same year of *i-Ready Diagnostic* only (comparison) or *Supplemental Blended* implementation. Baseline equivalence was met based on WWC guidance of an effect size with an absolute value no larger than 0.25, using a nearest neighbor matching approach.

Table 1 provides a summary of final baseline equivalence results, including the intra-class correlations (ICCs) that describe within group variance on *i-Ready Diagnostic* scores, total number of schools and students in the final sample, *i-Ready Diagnostic* mean and standard deviation, mean difference between the comparison and treatment groups, and hedge's *g* effect size.

Table 1. Baseline Equivalence Statistics for Matched Supplemental Blended (Treatment) Sample to i-Ready Diagnostic Only (Comparison)

Group	ICC	Schools	Students	*i-Ready Diagnostic* Mean	*i-Ready Diagnostic* SD	Adj Mean Diff	Effect Size
i-Ready Diagnostic only (Comparison)	0.16	31	847	473.80	31.27		
Supplemental Blended (Treatment)	0.28	35	854	468.56	32.38	-5.24	-0.165

Impact Analysis

Following the selection of baseline equivalent groups, hierarchical linear modeling (HLM) was used to address our research question to estimate the impact of *Supplemental Blended* implementation on student mathematics achievement. A two-level model was used to account for the clustered nature of the data with students nested within schools. Students identified during matching were those included in the impact analysis. Because effect size differences between the treatment and comparison at baseline fell between 0.05 and 0.25 standard deviations, statistical adjustment to account for baseline reading and mathematics achievement was incorporated into the model.

Level 1 of the model was specified as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(PRE\_MATH_{ij} - PRE\_MATH_{.j})_{ij} + \beta_{2j}(PRE\_READ_{ij} - PRE\_READ_{.j})_{ij} + \Sigma\beta_{qj}(GRADE)_{ij} + e_{ij}$$

where Y_{ij} is the student outcome for student *i* in school *j*. β_{0j} is the adjusted mean outcome for students in school *j*. β_{1j} and β_{2j} are the adjusted differences in outcome due to the student's pretest score in mathematics and reading, cluster (i.e., school) mean centered. β_{qj} is a vector of K minus 1 dummy variables to account for student grade level. e_{ij} is the random error in the achievement outcome associated with student *i* in school *j*.

Level 2 of the model was specified as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(GROUP)_j + \gamma_{02}(PRE\_MATH_{.j} - PRE\_MATH..)_j + \gamma_{03}(PRE\_READ_{.j} - PRE\_READ..)_j + \gamma_{04}(COHORT2)_j + \gamma_{05}(COHORT3)_j + \Sigma\gamma_K(SCHOOL) + u_{0j}$$

where γ_{00} is the adjusted control group grand mean of the outcome, γ_{01} is the adjusted mean difference in the outcome between school study groups, and GROUP is an indicator variable coded as 1 for schools in the *Supplemental Blended* group and 0 for schools in the *i-Ready Diagnostic* only comparison group. γ_{02} and γ_{02} are the regression slopes of the school-level pretests in mathematics and reading (grand mean centered) to explain additional between-school variance not explained in level 1 of the model. γ_{04} is the adjusted difference in the outcome for schools in Cohort 2, and γ_{05} is the adjusted difference in the outcome for schools in Cohort 3. γ_K is a vector of school-level demographic dummy variables added to increase statistical precision. u_{0j} is the random error in the achievement outcome associated with school *j*.

The student-level covariates used in each analysis were:

- Reading and mathematics *i-Ready Diagnostic* baseline performance
- Grade level

The school-level covariates included:

- Group membership (*Supplemental Blended* or *i-Ready Diagnostic* only)
- *i-Ready Diagnostic* average baseline performance on reading and mathematics
- Cohort (when treatment was started)
- Demographic variables[1]
  - Total school enrollment
  - Student-teacher ratio
  - Percent white
  - Percent students identified by Individuals with Disability Act (IDEA)
  - Percent students Limited English Proficient (LEP)
  - Title 1 distinction (yes or no)

The above variables were selected for inclusion in the model because they are exogenous and could reasonably be expected to be related to student achievement.

To indicate the size of impacts, effect sizes were computed for all comparisons using Hedge's *g*.

Impact Analysis Results

This section describes the results of the HLM analysis. Full information on the HLM model results, including student- and school- level covariate parameters, can be found in Appendix B. Table 2 summarizes the impact analysis findings. As shown, the *Supplemental Blended* schools were found to perform significantly better than the *i-Ready Diagnostic* only schools on mathematics achievement as measured by the *i-Ready Diagnostic*, with those in the treatment group on average performing almost 8 points higher than the comparison. The effect size of this difference was 0.22, as measured by Hedge's *g*.

Table 2. Impact Analysis Results for Supplemental Blended (Treatment) Schools Compared to i-Ready Diagnostic Only (Comparison) Schools for Mathematics at grades 6–8

Group	ICC	Schools	Students	i-Ready Mean	i-Ready SD	Adj Mean Diff	SE	P-Value	Effect Size
i-Ready only	0.15	31	847	487.43	33.00				
Supplemental Blended	0.34	35	854	495.27	38.48	7.84	2.14	<0.001	0.22

* = statistically significant

[1]School-level demographic variables obtained from the National Center for Education Statistics https://nces.ed.gov/ccd/

Impact Summary and Discussion

The findings of this impact study provide evidence that school-level implementation of *Supplemental Blended* resulted in increased student achievement in mathematics at grades 6–8 over use of *i-Ready Diagnostic* only. Schools using *Supplemental Blended* for mathematics showed significantly higher student achievement in mathematics compared to schools using *i-Ready Diagnostic* only for grades 6–8. The effect size of the difference was 0.22. For studies of educational intervention, it has been found that effect sizes of 0.25 or larger can be considered large (Lipsey et al., 2012). The effect size in our study was only slightly less than this threshold.

There are various limitations to this study. First, *i-Ready Diagnostic*, *Supplemental Ready* curriculum, and *i-Ready Instruction* were carried out in real-world conditions – the only requirements for inclusion in our treatment sample was that schools had purchased *Ready* curriculum products and students had at least some use of *i-Ready Instruction* in addition *to i-Ready Diagnostic* scores. Schools in both study groups were not participants in a research study but actual customers and everyday users of the products. Implementation thus varied between schools and within study groups. Impacts are typically greater for studies that aim for ideal or close to ideal implementation and less for studies that examine real-world implementation. Despite this limitation, we found a statistically significant effect for schools using the program under typical conditions.

Additionally, for this study, we used *i-Ready Diagnostic* only schools that have not used *Ready Supplemental* or *Ready Core* curriculum as our comparison group; we did not verify the core curricula used by these schools and, thus that information was not factored into the findings. If the curricula used by some of these schools were similar to *Ready* in content and rigor, our chances of detecting impacts may have been reduced. Future studies may benefit from further investigation into the curriculum use of comparison schools.

Finally, our treatment group was compared to an *i-Ready Diagnostic* only group. It is possible that use of *i-Ready Diagnostic* only increases student achievement; however, the design of this study did not allow for an estimation of that impact. Further, use of the *i-Ready Diagnostic* only as a comparison group may have attenuated the effects of the treatment group had that group been compared to a “business as usual” comparison group.

Quality Control Procedures

We employed various quality control checks throughout the analysis process. HumRRO, Curriculum Associates, and Century Analytics worked together to identify a rigorous methodology based on available data, the WWC 4.0 standards, and ESSA Level 2 guidelines. Data analysis work was completed collaboratively by HumRRO and Century Analytics. Century Analytics provided baseline equivalence and impact analysis data sets and output files. HumRRO confirmed only eligible schools and students participated and checked the output files and effect size calculations for accuracy. HumRRO generated final tables and figures based on the final baseline data and impact analysis output files; these tables were then checked by Century Analytics. HumRRO and Century Analytics were in frequent communication during the analyses to review the data and ensure accuracy of values and interpretations.

References

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at http://ies.ed.gov/ncser/.

U.S. Department of Education, Office of Elementary and Secondary Education. (2016, September). Non-Regulatory Guidance: *Using Evidence to Strengthen Education Investments.* Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2017, October). *What Works Clearinghouse: Standards Handbook (Version 4.0).* Retrieved from http://whatworks.ed.gov

Appendix A.

i-Ready Diagnostic Reliability Information from Curriculum Associates' Technical Report

Chapter 4. Reliability

Test reliability generally refers to the precision with which measurements are made (Haertel, 2006). All psychometric models rely on the notion of an underlying "true score" for each student which is measured imperfectly (i.e., with error) by any test. Analyses of test reliability endeavor to quantify the relative magnitude of true score and error variance that exist in observed test scores. The greater the proportion of true score variance relative to error variance in the observed variance of test scores, the more reliable the test. This proportional relationship exposes the indirect relationship between test reliability and measurement error: more reliable tests tend to give rise to scores with less measurement error. Because one cannot directly observe either the true score or error, a variety of methods are employed to predict measurement error and analyze test reliability. This chapter presents the reliability estimates for the Diagnostic and Growth Monitoring.

4.1. i-Ready Diagnostic

Three ways of characterizing test reliability for the i-Ready Diagnostic are included. First, because the i-Ready Diagnostic is an IRT-based CAT assessment, a student's standard error of measurement (SEM) may be calculated directly from students' item responses and the item difficulty parameters. In addition, the SEM varies with students' scores, so the conditional standard error of measurement (CSEM) is also presented. More reliable tests will give rise to lower SEMs.

Second, because the i-Ready Diagnostic adapts to students' item responses and delivers items that are targeted to their exhibited proficiency, a slightly different approach to estimating reliability is called for than for a fixed-form assessment. Classical test theory statistics like coefficient α assume uniform error variance across the scale. Under the IRT model on which the i-Ready Diagnostic is based, the standard error of measurement has a well-known and variable relationship with students' scale scores. As such, a more appropriate method for estimating reliability for a CAT assessment like the i-Ready Diagnostic is to calculate marginal reliability (Sireci, Thissen, & Wainer, 1991). Marginal reliability is more closely conceptually related to the definition of reliability based on the standard error of estimation given by Lord and Novick's (1968) Expression 3.8.4 than to internal consistency measures like coefficient α (Sireci, Thissen, & Wainer, 1991). In short, measurement error is calculated for each student and then the marginal error variance is calculated across all scale score levels, thus appropriately accounting for the fact that scale scores toward the middle of the scale tend to be more precise than at either extreme. More reliable tests will have greater marginal reliability estimates.

Third, test-retest reliability analyses are presented, in which students test twice and the strength of the linear relationship between their scores is yet another lens through which to view reliability. More reliable test scores are strongly, linearly related, because they contain relatively little error variance (when compared with true score variance).

4.1.1. Standard Error of Measurement (SEM)

The Standard Error of Measurement (SEM) is a measure of the degree of precision of students' i-Ready Diagnostic scores. SEMs are affected by factors such as how well the data fit the underlying model, student response consistency, student location on the ability continuum, match of items to student ability, and test length. In the context of i-Ready assessments, a high SEM could be caused by students performing erratically or having extreme response vectors (e.g., getting all items correct or incorrect). Although there are no specific targets for observed standard errors, lower values of standard errors are preferable to higher values because they suggest reliable student ability measures. Given the adaptive nature of i-Ready and the wide difficulty range in the item bank, standard errors are expected to approach the theoretical minimum for the test of the given length. The theoretical minimum would be reached if each item difficulty value matched the student's ability estimate perfectly. Theoretical minimums are restricted by the number of items served in the assessment—the more items that are served up, the lower the SEM could potentially be.

The theoretical minimum SEM for the scale score is given by the following equation:

$$Min(SEM) = \frac{b}{\sqrt{n \times 0.25}} \quad (5)$$

where b is the scaling constant which is the slope for scale transformation from θ to the i-Ready scale (more detail about scale scores is provided in section 7.1.1), n is the number of items, 0.25 is the expected maximum information (hence minimum error) under the Rasch model (when the probability of a correct and of an incorrect response are both equal to 0.5). For the overall score, a typical assessment consists of 72 items. The number of items within a domain varies between 12 and 36 depending on the test flow (see Appendices D and E for more detail about the test flow). The theoretical minimum SEM for overall scores is 8.9 and 6.0 scale

score points for Reading and Mathematics, respectively. The theoretical minimum SEMs for domain scores range from 17.8 to 21.8 for Reading (12 to 18 items per domain) and range from 8.5 to 13.6 for Mathematics (14 to 36 items per domain).

Table 4.1 shows the mean and standard deviation of the overall score SEM across all Diagnostic assessments taken during August and September of the 2016–2017 school year[7]. The mean SEMs for overall scores are low in both the Reading (9.3–10.9) and Mathematics assessments (6.3–6.5), with many approaching the theoretical minimum SEM.

Table 4.1. August and September of 2016–2017 School Year: Diagnostic Mean SEMs for Overall Scores and Sample Size by Grade and Subject

	Reading												
Statistic	K	1	2	3	4	5	6	7	8	9	10	11	12
Mean	9.3	9.3	10.4	10.0	10.1	10.3	10.5	10.6	10.7	10.7	10.8	10.9	10.8
N*	184.3	287.6	323.3	366.3	346.0	345.2	251.3	225.9	223.6	33.9	22.3	14.6	7.4
	Mathematics												
Statistic	K	1	2	3	4	5	6	7	8	9	10	11	12
Mean	6.5	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.3	6.3	6.3	6.4
N*	191.2	298.5	334.2	376.1	366.0	366.1	276.3	254.2	238.8	39.5	25.4	16.4	8.5

*: Sample size in thousands of students.

Table 4.2 shows the mean SEMs by domain in the Diagnostic tests. Each domain is assessed by 12–36 items (see Appendices B & C for details about the number of items per domain). The observed mean SEMs are also close to the respective minimal value given the length of the domain.

Table 4.2. August and September of 2016–2017 School Year: Diagnostic Mean SEMs for Domain Scores by Grade

Reading Domain	K	1	2	3	4	5	6	7	8	9	10	11	12
Comprehension: Informational Text	24.3	24.1	24.1	19.2	19.4	19.5	19.6	19.7	19.8	19.9	20.0	20.2	20.3
Comprehension: Literature	24.2	24.2	24.5	19.4	19.5	19.6	19.7	19.8	19.9	20.1	20.2	20.4	20.4
High-Frequency Words	24.4	26.9	36.9	28.3	29.0	29.1	28.8	28.5	28.5	33.4	34.1	34.5	30.4
Phonological Awareness	24.1	24.6	24.4	24.8	24.7	24.4	24.9	24.8	24.3	25.3	24.7	25.1	23.8
Phonics	23.6	23.6	23.5	24.5	24.6	24.8	24.8	24.9	25.0	25.0	25.2	25.1	25.3
Vocabulary	25.6	24.0	23.9	18.8	18.8	18.7	18.7	18.8	18.8	18.9	19.0	19.1	19.4
Mathematics Domain	K	1	2	3	4	5	6	7	8	9	10	11	12
Algebra and Algebraic Thinking	12.9	12.6	12.6	12.5	12.6	12.6	12.6	12.6	12.6	10.3	10.2	10.1	10.6
Geometry	14.8	14.9	14.7	14.5	14.5	14.5	14.5	14.5	14.5	11.0	10.9	10.5	11.0
Measurement and Data	14.8	14.7	14.5	14.6	14.6	14.6	14.6	14.6	14.6	14.8	14.8	14.8	14.9
Number and Operations	11.9	11.9	11.9	12.0	11.9	11.9	11.9	11.9	11.9	12.1	12.1	12.1	12.2

Table 4.3 shows the mean overall score SEMs across Diagnostic assessments taken by three special groups of students for August and September of the 2016–2017 school year: Economically Disadvantaged, English Learner (EL), and Special Education students. The mean SEMs for these special groups are low and very similar to the mean SEMs for the entire i-Ready population presented in Table 4.1.

[7] For SEM and CSEM based on the 2015–2016 data based on the previous scale, please refer to Appendix M.

Table 4.3. August and September of 2016–2017 School Year: Diagnostic Mean SEMs for Overall Scores and Sample Size by Grade and Subject for Special Groups

	Reading													
Group	N	K	1	2	3	4	5	6	7	8	9	10	11	12
Economically Disadvantaged	258.7	9.3	9.3	10.2	9.9	10.0	10.2	10.4	10.5	10.6	10.7	10.8	10.8	10.7
EL	237.9	9.3	9.3	10.1	9.7	9.7	9.9	10.0	10.2	10.3	10.6	10.6	10.7	10.6
Special Ed.	158.4	9.3	9.4	10.0	9.7	9.8	10.0	10.0	10.2	10.3	10.4	10.5	10.6	10.6
	Mathematics													
Group	N	K	1	2	3	4	5	6	7	8	9	10	11	12
Economically Disadvantaged	333.7	6.5	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.3	6.3	6.4	6.4
EL	255.2	6.5	6.4	6.4	6.4	6.4	6.4	6.4	6.5	6.5	6.4	6.4	6.4	6.4
Special Ed.	172.7	6.5	6.5	6.5	6.4	6.5	6.4	6.5	6.5	6.5	6.5	6.5	6.5	6.6

4.1.2. Conditional Standard Error of Measurement

In addition to the mean SEMs for both overall and domain scores, Figures 9 and 10 present the observed mean conditional SEMs by overall score across the full spectrum of the i-Ready scale based on Diagnostic assessments administered in August and September of the 2016–2017 school year. The middle 98% range of the observed scores is indicated for both subjects. Except at the very tail ends of the scale, more than 98% of the score ranges have conditional SEMs around 10 (logit value of 0.26) for Reading, and lower than 7.5 (logit value of 0.30) for Mathematics. The results shown in these figures demonstrate that the measurement error of the i-Ready Diagnostic is consistent and at a desirable (i.e., low) level across a wide score range. Considering these values relative to the theoretical SEM minima of 8.9 and 6.0 scale score points for Reading and Mathematics, respectively, reveals that the i-Ready Diagnostic scale scores exhibit very low measurement error and therefore a high degree of reliability.

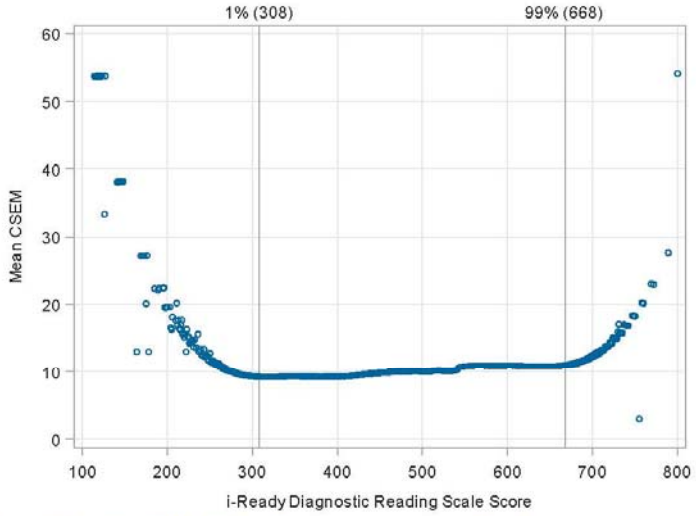

Figure 9. Conditional SEM of i-Ready Diagnostic Assessments for Reading

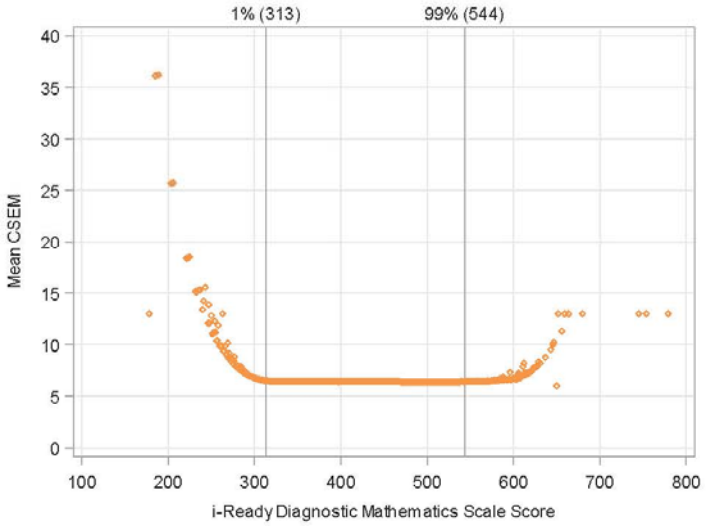

Figure 10. Conditional SEM of i-Ready Diagnostic Assessments for Mathematics

4.1.3. Marginal Reliability Estimates

The IRT analogue to classical reliability estimation is called marginal reliability and operates on the variance of the theta scores and the mean of the expected error variance (Samejima, 1977; Sireci, Thissen, & Wainer, 1991). This marginal reliability uses the classical definition of reliability as proportion of variance in the total observed score due to true score. The true score variance is computed as the observed score variance minus the error variance (see equation below).

$$\rho_{\theta} = \frac{\sigma^2_{\hat{\theta}} - \bar{\sigma}^2_{\hat{E}}}{\sigma^2_{\hat{\theta}}} \quad (6)$$

where ρ_{θ} is the marginal reliability estimate, $\sigma^2_{\hat{\theta}}$ is the observed error variance of the ability estimate, and $\bar{\sigma}^2_{\hat{E}}$ is the observed mean conditional error variance. Like a classical reliability coefficient, the marginal reliability estimate increases as the standard error of measurement decreases; it approaches 1 when the standard error of measurement approaches 0.

Other relevant measures provided by WINSTEPS *(Linacre, 2012; Linacre, 2002)* in Rasch analysis are separation indices and reliability estimates. Separation indices reflect the ratio of person (or item) standard deviation to the standard deviation of error and are closely related to reliability estimates *(Wright, 1996)*. Values in excess of 2.0 indicate that greater than 80% of the variance in scores is not due to error, but rather to person or item differences. For the more common reliability measures, person reliability (or marginal reliability) is equivalent to the commonly recognized test score reliability in classical test theory settings such as the KR–20 internal consistency reliability coefficient. Further, Rasch analysis provides item reliability or the ratio of true item variance to observed item variance, which has no direct counterpart in classical test analysis. The higher item reliability values indicate greater consistency of item ordering.

Based on the spring 2014 calibration, the estimated reliability for the Reading test was 0.97 with a separation index of 5.43 and the estimated reliability of the Mathematics test was 0.96 with a separation index of 5.22. In addition, data from August and September 2016 were used to estimate marginal reliability for each domain score by grade (Table 4.5) and for overall scores by grade (Table 4.4). Table 4.5 shows that—apart from Grade 2 High-Frequency Words—reliability estimates by domain and grade vary between 0.63 and 0.96. The marginal reliability estimates for some domains were lower than others (e.g., High-Frequency Words and Phonological Awareness), likely due to the shorter test length as well as students' grade restriction related to the content for these

domains. Since the overall score is based on all domains and hence a greater number of items, Table 4.4 shows that the overall scores tend to be much more reliable than any individual component domain, with marginal reliability estimates of between 0.91 and 0.99.

Table 4.4. August and September of 2016–2017 School Year: Marginal Reliability Estimates for Overall Scores by Grade

	Reading												
Statistic	K	1	2	3	4	5	6	7	8	9	10	11	12
Marginal Reliability	0.91	0.95	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98
N*	184.3	287.6	323.3	366.3	346.0	345.2	251.3	225.9	223.6	33.9	22.3	14.6	7.4
	Mathematics												
Statistic	K	1	2	3	4	5	6	7	8	9	10	11	12
Marginal Reliability	0.92	0.94	0.94	0.95	0.96	0.96	0.96	0.97	0.97	0.98	0.98	0.98	0.99
N*	191.2	298.5	334.1	376.0	365.7	365.9	276.2	254.0	238.6	39.5	25.4	16.4	8.5

*: Sample size in thousands of students.

Table 4.5. August and September of 2016–2017 School Year: Marginal Reliability Estimates for Domain Scores by Grade

Reading Domain	K	1	2	3	4	5	6	7	8	9	10	11	12
Comprehension: Informational Text	0.70	0.79	0.85	0.90	0.91	0.91	0.91	0.92	0.92	0.93	0.94	0.94	0.95
Comprehension: Literature	0.68	0.77	0.84	0.89	0.90	0.90	0.91	0.91	0.91	0.92	0.93	0.93	0.95
High-Frequency Words	0.64	0.76	0.57	0.75	0.76	0.76	0.76	0.76	0.75	0.78	0.74	0.81	0.80
Phonological Awareness	0.66	0.76	0.73	0.78	0.79	0.77	0.83	0.81	0.82	0.86	0.81	0.83	0.80
Phonics	0.71	0.80	0.84	0.79	0.80	0.79	0.79	0.79	0.80	0.81	0.82	0.83	0.87
Vocabulary	0.63	0.80	0.82	0.88	0.88	0.89	0.89	0.90	0.90	0.91	0.92	0.93	0.95
Mathematics Domain	K	1	2	3	4	5	6	7	8	9	10	11	12
Algebra and Algebraic Thinking	0.75	0.81	0.82	0.84	0.85	0.86	0.87	0.88	0.89	0.92	0.93	0.94	0.96
Geometry	0.78	0.79	0.81	0.81	0.83	0.85	0.86	0.87	0.88	0.92	0.93	0.94	0.95
Measurement and Data	0.74	0.79	0.81	0.83	0.84	0.86	0.87	0.87	0.88	0.86	0.87	0.87	0.91
Number and Operations	0.76	0.80	0.81	0.82	0.85	0.87	0.89	0.90	0.91	0.84	0.84	0.85	0.89

4.1.4. Test-retest Reliability Estimates

Evidence of test-retest reliability was assessed based on a subsample of students who, during the 2016–2017 school year, took the i-Ready Diagnostic once in the fall and once in the winter testing windows with the recommended 12–18 weeks between tests and rushed in neither test. The mean testing interval was 106 days (15.2 weeks), and over that period students would have received classroom instruction and hence have been anticipated to have grown. The fact that the test administrations were more than a negligible amount of time apart suggests that not only was there variability in the test scores due to error variance, but also that students true scores may have changed and therefore introduced additional variance that is not attributable to the precision of the i-Ready Diagnostic test itself. Table 4.6 presents the correlations between the two overall Diagnostic scores. For all grades and both subjects, test-retest correlations are above 0.70, which, while considered strong, tends to be lower than the marginal reliability estimates, primarily because of the likelihood that students' content knowledge did grow over the 12–18 weeks that passed between their two test occasions.

Table 4.6. Test-Retest Correlations by Grade and Subject

	Reading												
Statistic	K	1	2	3	4	5	6	7	8	9	10	11	12
Test-Retest Reliability	0.70	0.82	0.85	0.85	0.86	0.86	0.86	0.85	0.85	0.86	0.86	0.85	0.88
N*	118.1	159.1	173.3	199.5	193.5	190.3	137.1	119.9	114.4	12.5	8.2	4.9	2.0
	Mathematics												
Statistic	K	1	2	3	4	5	6	7	8	9	10	11	12
Test-Retest Reliability	0.71	0.77	0.81	0.82	0.85	0.86	0.87	0.87	0.87	0.85	0.85	0.86	0.89
N*	113.4	161.9	184.4	210.8	210.2	206.9	153.9	134.1	123.3	13.0	7.0	3.7	1.8

*: Sample size in thousands of students.

4.2. Standard Error of Measurement for Growth Monitoring

Table 4.7 shows the mean SEMs for Growth Monitoring (GM) assessments taken during August and September of the 2016–2017 school year[8]. Because the Growth Monitoring tests are much shorter (19–21 items), the theoretical minimum SEM is between 16.5 and 17.3 points for Reading and 11.1 and 11.7 points for Mathematics. The observed SEMs are close to the lowest obtainable value.

Table 4.7. August and September of 2016–2017 School Year: Growth Monitoring SEM and Sample Size by Grade

	Reading								
Statistic	K	1	2	3	4	5	6	7	8
Mean	17.7	17.7	17.7	17.8	17.8	17.8	17.9	17.9	17.9
N*	36.5	63.0	70.0	78.5	69.0	65.2	29.2	20.3	17.4
	Mathematics								
Statistic	K	1	2	3	4	5	6	7	8
Mean	11.8	11.8	11.8	11.8	11.8	11.8	11.8	11.8	11.9
N*	26.3	46.9	54.4	65.3	62.6	60.1	35.3	26.8	22.5

*: Sample size in thousands of students.

[8] For SEM and CSEM based on the 2015–2016 data based on the previous scale, please refer to Appendix M.

Appendix B. Impact HLM Coefficients

Table B.1. HLM Results for Supplemental Blended Compared to i-Ready Diagnostic Only for Grades 6–8

Covariates	Coef.	SE	z	p-value	95% Conf. Interval	
Treatment Group Membership	7.842	2.138	3.67	0.000	3.651	12.032
Fall Reading *i-Ready* School-Level Grand Mean Centered	0.073	0.065	1.13	0.260	-0.054	0.200
Fall Math *i-Ready* School-Level Grand Mean Centered	0.878	0.129	6.8	0.000	0.625	1.130
Fall *i-Ready* Reading Student-Level School Mean Centered	0.120	0.013	9.01	0.000	0.094	0.146
Fall *i-Ready* Math Student-Level School Mean Centered	0.740	0.022	33.18	0.000	0.696	0.784
Grade*						
7	0.209	1.206	0.17	0.862	-2.154	2.572
8	-1.625	1.209	-1.34	0.179	-3.995	0.745
Cohort**						
Cohort 2 (2015–16)	2.714	3.041	0.89	0.372	-3.246	8.674
Cohort 3 (2016–17)	1.594	2.491	0.64	0.522	-3.288	6.476
School-Level Covariates						
Total Enrollment	0.004	0.004	1.11	0.269	-0.003	0.012
Percent White	0.189	0.059	3.2	0.001	0.073	0.305
Percent LEP	0.006	0.133	0.05	0.962	-0.254	0.266
Percent IDEA	0.313	0.260	1.2	0.229	-0.197	0.823
Title 1 Designation	2.289	2.691	0.85	0.395	-2.984	7.563
Student-Teacher Ratio	0.111	0.290	0.38	0.702	-0.458	0.680
intercept	461.405	10.887	42.38	0.000	440.067	482.742

*Grade 7 and Grade 8 were included in the analysis model as covariates to adjust for student grade level and increase precision. Grade 6 was the reference grade.

**Cohort 2 and Cohort 3 were included in the analysis model as covariates to adjust for cohort and increase precision. Cohort 1 (2014–2015) was the reference cohort.