HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2019 No. 074

An Impact Evaluation of Reading *i-Ready Instruction* for Middle School Grades

Final Report

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

Authors:
Rebecca Norman Dvorak
Bruce Randel (Century Analytics)
Matthew Swain

Date: September 16, 2019

An Impact Evaluation of Reading *i-Ready Instruction* for Middle School Grades

Table of Contents

An Impact Evaluation of Reading *i-Ready Instruction* for Middle School Grades 3

Introduction 3
- Defining i-Ready Instruction 3

Research Questions 4

Methodology 4
- Design Decisions 4
  - Cluster-Level Design 4
  - Outcome Measure 4
  - Required Number of Schools 5
- School and Student Selection Criteria 5
  - i-Ready Instruction Schools and Students 6
  - Comparison Schools 8
- Matching 9
  - School-Level Matching 9
  - Student-Level Matching 11
- Impact Analysis 11
- Impact Analysis Results 12

Impact Summary and Discussion 15
- Quality Control Procedures 16

References 17

Appendix A. *i-Ready Instruction* Theory of Action A-1

Appendix B. i-Ready® Diagnostic Reliability Information from Curriculum Associates' Technical Report B-1

Appendix C. Impact HLM Coefficients C-1

Table of Contents (Continued)

List of Tables

Table 1. School Characteristics of i-Ready Instruction Middle Schools at Key Points in the Filtering Process. 8
Table 2. School Characteristics of Comparison Schools at Key Points in the Filtering Process.. 8
Table 3. School Demographic Variables and Effect Size Differences Between the i-Ready Instruction (Treatment) and Comparison Group. 10
Table 4. Urbanicity Differences Between the i-Ready Instruction (Treatment) and Comparison Group.....10
Table 5. Reading Baseline Equivalence Statistics for i-Ready Instruction (Treatment) and Comparison Groups by Grade 11
Table 5. Impact Analysis Results for i-Ready Diagnostic Schools Compared to Comparison Schools for Reading Student Achievement at grades 6–8.....14
Table C.1. HLM Results for Grade 6 Reading. C-1
Table C.2. HLM Results for Grade 7 Reading. C-2
Table C.3. HLM Results for Grade 8 Reading. C-2

An Impact Evaluation of Reading *i-Ready Instruction* for Middle School Grades

Introduction

Founded in 1969, Curriculum Associates provides a variety of educational products and services with the goal of improving education for students and teachers. Two Curriculum Associates products include *i-Ready® Diagnostic* (available for K–12) and *i-Ready® Instruction* (available for K–8). The *i-Ready Diagnostic* assessments (a) are online, computer-adaptive assessments that pinpoint student needs at the sub-skill level and (b) help monitor the extent to which students are on track to achieve end-of-year targets. *i-Ready Instruction* provides online, individualized instruction for students and can be used with the *i-Ready Diagnostic.*

The Human Resources Research Organization (HumRRO) conducted an evaluation to examine the impact of *i-Ready Instruction* on reading achievement for students in middle school grades 6–8. We designed this study 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, 2017a) 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). We achieved this by using quasi-experimental design (QED), establishing baseline equivalence between the treatment and comparison groups, using an outcome measure acceptable by WWC, including baseline achievement as a covariate, and using a sampling design that mitigates the effects of any confounding factors.

Defining i-Ready Instruction

i-Ready Instruction is an online personalized instruction program aligned to college- and career-ready standards that includes engaging multimedia instruction and progress monitoring into online lessons. Lessons are intended to provide a consistent best-practice lesson structure and build students' conceptual understanding. *i-Ready Instruction* is intended to be used in conjunction with *i-Ready Diagnostic* which monitors student progress and identifies student performance in reading. This diagnostic information helps target student-specific intervention, which is provided through *i-Ready Instruction.*

Curriculum Associates has developed a Theory of Action (TOA) noting the key implementation components of *i-Ready Instruction*, the intended intermediate outcomes, and the intended long-term outcomes. The key implementation components highlight actions by students, teachers, and leaders recommended to obtain the long-term outcome of improved student learning in reading. Among others, the key components include support at the school and district leadership level, monitoring of student progress by teachers, and student use of *i-Ready Instruction* to work through a personalized, scaffolded instruction path. The *i-Ready Instruction* TOA is provided in Appendix A.

Curriculum Associates provides guidance to districts and schools on how to implement *i-Ready Instruction* to best benefit student learning (Curriculum Associates, 2019). Guidance indicates students receive greater gains with at least 30 minutes of *i-Ready Instruction* use for each subject area. In addition, Curriculum Associates recommends use for at least 18 weeks between a fall *i-Ready Diagnostic* administration and a spring administration (Curriculum Associates, 2018).

Research Questions

The primary purpose of this evaluation was to estimate the impact of using *i-Ready Instruction* on student reading achievement. Our research was focused on one primary research question, addressed separately for each middle school grade 6 to grade 8.

- What was the impact of *i-Ready Instruction* use in reading on student reading achievement as compared to schools and students who do not use *i-Ready Instruction* in reading?

Our hypothesis was that student achievement for reading would be higher for students in schools that implement *i-Ready Instruction* with fidelity according to the criteria described in the TOA and user guidance (Curriculum Associates, 2019). This hypothesis was based on the belief that students benefit from the *i-Ready Instruction* targeted to their specific needs in reading.

Methodology

This section describes initial design decisions, school and student selection and matching, analysis, and results.

Design Decisions

Cluster-Level Design

Though it is most common for Curriculum Associates to sell a product, or set of products, at the district-level, usage data and on-the-ground experiences support the notion that the decision to use *i-Ready Instruction*, and to what degree, happens at the school-level. Therefore, we determined the unit of assignment for this study to be schools. The unit of observation in this study was the student and our analyses focused on the impact of *i-Ready Instruction* use on student achievement in reading for students and schools using *i-Ready Instruction* with fidelity.

Outcome Measure

The reading *i-Ready Diagnostic* assessments were designed to be aligned to today's college- and career-ready standards and to provide results that inform student placement decisions, offer explicit instructional advice, and prescribe resources for targeted instruction and intervention. The *i-Ready Diagnostic* is currently used by more than 6.5 million students across the United States – sometimes incorporating other *i-Ready* products (i.e., *Instruction*, Teacher Resources), though this is not a requirement.

To provide evidence the *i-Ready Diagnostic* measures skills consistent with student expectations, Curriculum Associates has conducted multiple linking studies to examine the correlation of *i-Ready Diagnostic* scores with scores from national and state summative tests for reading at grades 3 – 8. Linking studies using 2016 data examined the correlation between *i-Ready Diagnostic* and the Smarter Balanced summative assessments, the Partnership for Assessment of Readiness for College and Careers (PARCC), and multiple state testing programs (North Carolina, New York, Tennessee, Ohio, Mississippi, Michigan, Indiana, Florida, and Georgia). These studies show strong correlations between *i-Ready Diagnostic* scores and scores on these national and state tests. The average correlations across grades between the *i-Ready Diagnostic* for reading and the national and state reading assessments the correlations ranged from 0.78 (Tennessee TNReady) to 0.85 (Smarter Balanced). These studies provide

evidence that the *i-Ready Diagnostic* content is highly consistent with what students across the United States are expected to learn (Curriculum Associates, 2019a). Curriculum Associates has also recently completed linking studies for Colorado, Kentucky, and Missouri. In addition, Curriculum Associates has commissioned Odell Education and others to complete alignment studies to demonstrate the degree of alignment between the content on *i-Ready Diagnostic* and current sets of state standards. Specifically, they have conducted alignment studies for the Common Core State Standards (CCSS), and for the Louisiana, Indiana, Ohio, Michigan, Florida, and South Carolina state standards.

Curriculum Associates released *i-Ready* in the summer of 2011. Since then, *i-Ready* has been reviewed and approved at the state level as an assessment, instructional resource, or intervention in Arizona, California, Colorado, Connecticut, Delaware, Florida, Georgia, Idaho, Indiana, Massachusetts, Mississippi, Nevada, New Mexico, New York, North Carolina, Ohio, Oklahoma, Oregon, Tennessee, Utah, and Virginia.

i-Ready Diagnostic for reading measures achievement aligned to common reading content and skills with demonstrated test score reliability (see Appendix B). Marginal reliabilities for reading at grades 6 – 8 were each .97 and test-retest reliabilities ranged from .85 to .86. Therefore, this assessment meets the WWC 4.0 standards for an acceptable outcome measure (WWC, 2017a). The *i-Ready Diagnostic* was used as the baseline and outcome measure for all students participating in this study (i.e., *i-Ready Instruction* students and comparison group students).

The *i-Ready Diagnostic* is intended to be administered in a standardized manner across schools (Curriculum Associates, 2019b). Specifically, at grades 6–8 teachers are to schedule the first (fall) Diagnostic 2 – 3 weeks into the school year in two 45- to 50-minute sessions. They are to test technology to ensure proper function and have pencils and paper available as scratch paper. Test administrators are to provide instructions to their students and motivate them to do their best. As students are testing, teachers are to monitor students.

Required Number of Schools

We conducted power analyses using Optimal Design software (Spybrook, et al., 2011) to identify the total number of schools required at each grade level for sufficient statistical power to reject the null hypothesis that there is no difference in student reading achievement between the treatment and comparison group. Statistical power is influenced by various factors. We used data from previous studies HumRRO conducted using *i-Ready Diagnostic* as an outcome to estimate conservative and optimistic parameters for use in the power analysis. These parameters were: (a) 0.90 for the relationship between the baseline and outcome variable, (b) 40 and 60 for the number of students per school, and (c) 0.10 and 0.30 for the intraclass correlation coefficient (ICC). Optimal Design estimated that approximately 40 schools would be needed to achieve statistical power of 0.80. A 0.80 power level provides an 80% chance of detecting a statistically significant difference if one exists with 95% confidence.

School and Student Selection Criteria

This section provides an overview of the process of selecting schools and students eligible for inclusion in the study. To make eligibility determinations, HumRRO used a combination of publicly available data and Curriculum Associates *i-Ready Diagnostic* and *i-Ready Instruction* usage data. Data provided by MDR[1] summarized recent publicly available school-level

[1] https://mdreducation.com/our-story/

enrollment data, free and reduced prince lunch participation data, and urbanicity data compiled from the Department of Education Common Core of Data and other sources. We used data available through the Department of Education Civil Rights Data Collection Website[2] to obtain the most recent publicly available school-level information on percent white, percent limited English proficiency (LEP), and percent students with disabilities (SWD).

HumRRO started with Curriculum Associates *i-Ready* 2017-18 datasets for reading. Because the *i-Ready Diagnostic* was selected as our student baseline and outcome measure for the treatment and comparison group, we included only students with both fall 2017 and spring 2018 *i-Ready Diagnostic* scores for reading. For the purpose of generalizability and interpretability, we identified middle schools with a similar set of characteristics. Particularly, we filtered our file to include only schools meeting the following criteria: (a) public institutions, (b) not charter or magnet schools, and (c) grades 6–8 only institutions. Restricting our sample in this manner helped reduce potential confounds due to institution type.

We next sought to identify schools qualified for inclusion in either the treatment or comparison groups. We separated out those schools with at least some *i-Ready Instruction* use (potential treatment) and those including only students with *i-Ready Diagnostic* use (potential comparison). Additional filtering was required for both groups to ensure we included schools meeting minimum fidelity requirements. These additional steps are described separately for each group.

i-Ready Instruction Schools and Students

One of our primary goals for this study was to include schools that implemented *i-Ready Instruction* for reading with fidelity. Specifically, we wanted to ensure that students in our schools were using *i-Ready Instruction* in a manner consistent with the logic model and the guidance provided by Curriculum Associates to its users. HumRRO, Century Analytics, and Curriculum Associates worked collaboratively to establish the definition of implementing *i-Ready Instruction* with fidelity at the student-level. First, based on current *i-Ready Diagnostic* guidance Curriculum Associates provides to its users, we determined a student would need to use *i-Ready Instruction* for reading for an average of 30 minutes or more per week of use during the 2017-18 school year. Second, a student would need to use *i-Ready Instruction* for at least 18 weeks between the fall and spring *i-Ready Diagnostic* administrations to be considered using with fidelity. HumRRO filtered the *i-Ready Instruction* dataset to include only schools with students meeting this definition.

Curriculum Associates staff indicated most schools were not fully implementing *i-Ready Instruction* until the second school-year of use. They found it generally took some time for teacher buy-in of the new intervention and to find ways to carve out the time recommendation for student use. Therefore, we determined it was appropriate to include schools in their second year of *i-Ready Instruction* use in our study to achieve our goal of examining the impact of use with fidelity. We combined 2015-16 and 2016-17 *i-Ready Diagnostic* and *Instruction* usage data with our 2017-18 data to make this determination. Schools were kept in our *i-Ready Instruction* school list if (a) they did not have any *i-Ready Instruction* for reading use in 2015-16, (b) they had at least some *i-Ready Instruction* for reading use in 2016-17, and (c) included students who met *i-Ready Instruction* fidelity criteria for reading during the 2017-18 school year. There were 209 schools that met this requirement.

[2] https://www2.ed.gov/about/offices/list/ocr/data.html

For a middle school to be using with fidelity, we determined at least 30% of students should be meeting the student-level *i-Ready Instruction* usage criteria. This percentage was determined after reviewing data and discussions between HumRRO, Curriculum Associates, and Century Analytics surrounding the nature of implementation in middle schools. We acknowledge it is common for middle school students to be split into different ELA courses, and some teachers may make different decisions regarding *i-Ready Instruction* use. Our data indicated, in general, a smaller percentage of students per school used *i-Ready Instruction* at an implementing middle school compared to what is typical in an elementary school. Using publicly available enrollment data, we calculated the approximate percentage of students per grade using *i-Ready Instruction* with fidelity. We next calculated the average percentage of students meeting the fidelity criteria across the three grades. Schools that had an average of 30% of students or more across all grades meeting the fidelity criteria were identified as those meeting all implementation fidelity criteria. This left us with 31 eligible schools.

As a final step, we sought confirmation from Curriculum Associates staff who work directly with the identified schools to ensure that (a) they were, in fact, using *i-Ready Instruction* with fidelity as our data suggested, and (b) that these schools were using *i-Ready Instruction* with their general population of students. Use could be confined to a particular English language arts (ELA) class or set of ELA classes but should not be used with only students identified for intervention and/or special education students. Based on this definition of implementation fidelity we removed two schools from the list of eligible *i-Ready Instruction* schools. In addition, we identified three schools for which we did not have complete data for all four matching variables. These schools were also removed. This resulted in 26 eligible schools for matching.

Table 1 presents key school characteristics for *i-Ready Instruction* schools at key points in the filtering process described above. Step 1 includes all public 6–8 schools in their second year of *i-Ready Instruction* implementation who have at least one student meeting minimum usage requirements for *i-Ready Instruction*. Step 2 includes only those schools with, on average, 30% students or more eligible for inclusion across grades. Finally, Step 3 summarizes data of our final set of schools confirmed by Curriculum Associates to be using *i-Ready Instruction* with a general education population of students. As shown, the student demographic make-up of students changes for some variables between Step 1 and the final set of schools meeting all school- and student-level eligibility criteria for the *i-Ready Instruction* group. Particularly, the middle schools meeting eligibility criteria have a higher percentage of white students and lower percentage of limited English proficient students compared to the unfiltered set of schools with at least some students meeting the *i-Ready Instruction* criteria. There is also a smaller percentage of urban schools in our final sample.

Table 1. School Characteristics of i-Ready Instruction Middle Schools at Key Points in the Filtering Process.

Variable	Step 1. School-Level Second Year *i-Ready Instruction* Users Criteria Applied (*n* = 209)	Step 2. 30% Minimum *i-Ready Instruction* Student Use Criteria Applied (n = 31)	Step 3. Final Approved for Matching (n = 26)
Percent FRL	58.24	61.46	57.52
Percent LEP	7.22	4.53	3.50
Percent SWD	15.96	16.86	16.98
Percent White	48.41	60.97	67.62
Urbanicity			
Rural	16.23	26.67	24.00
Suburban	41.58	33.33	40.00
Town	15.35	26.67	28.00
Urban	26.73	13.33	8.00

Comparison Schools

Filtering comparison schools began by removing schools with past exposure to *i-Ready Instruction*. HumRRO used *i-Ready Instruction* data from the two school years prior to the study year to identify and remove schools with any use of *i-Ready Instruction* for 2015-16 and 2016-17. To make sure our comparison sample was a pure comparison, we removed schools with as few as one student who had used *i-Ready Instruction* for reading for either of those two years, as indicated by the usage data. Following this filter, our comparison dataset included 58 schools having *i-Ready Diagnostic* data for fall and spring reading.

Next, following similar guidelines to our *i-Ready Instruction* sample, we identified schools where, at minimum, an average of 30% of students across grades 6–8 were administered the *i-Ready Diagnostic* in the spring and fall of 2017-18. This filter resulted in 34 eligible schools. Finally, consistent with the approach taken for *i-Ready Instruction* schools, we provided our filtered list of comparison schools to the Curriculum Associates staff who work directly with the schools to confirm our schools used *i-Ready Diagnostic* with their general student population. We learned that one school did not meet our definition of use with general education students. We removed this school from our eligible comparison dataset. This left 33 schools meeting eligibility requirements for inclusion as comparisons.

Table 2 presents school characteristics for comparison schools at key points in the filtering process described above. Step 1 includes all schools identified as public, grade 6–8 institutions with *i-Ready Diagnostic* users that had not even one *i-Ready Instruction* student user in 2017–18, 2016–17, or 2015–16. Step 2 includes only those schools with, at minimum, an average of 30% students with available *i-Ready Diagnostic* scores for mathematics across grades. Step 3 shows the schools that were confirmed by Curriculum Associates staff as using *i-Ready Instruction* with their general education students and met all comparison eligibility criteria. As shown, the schools eligible for inclusion in the final comparison group (Step 3) are of mostly similar demographic and urbanicity make-up to those in Step 1 prior to applying final usage

filters. One exception was that a smaller percentage of towns were included in our final matching sample compared to the sample prior to applying filters.

Table 2. School Characteristics of Comparison Schools at Key Points in the Filtering Process.

Variable	Step 1. School-Level Previous *i-Ready Instruction* Exposure Criteria Applied (n = 58)	Step 2. 30% Minimum *i-Ready Diagnostic* Student Use Criteria (n = 34)	Step 3. Final Approved for Matching (n = 33)
Percent FRL	46.51	48.48	48.47
Percent LEP	5.09	4.95	5.04
Percent SWD	16.81	16.75	16.75
Percent White	69.28	72.74	72.48
Urbanicity			
Rural	16.67	14.71	15.15
Suburban	42.59	47.06	48.48
Town	12.96	8.82	6.06
Urban	27.78	29.41	30.30

Matching

We conducted a two-step matching procedure. First, we matched *i-Ready Instruction* and comparison schools to be similar on school-level variables. Then, we examined means on the baseline measure between the students from the matched *i-Ready Instruction* and comparison schools.

School-Level Matching

Once we had identified eligible *i-Ready Instruction* and comparison schools, described above, we matched schools to ensure key school-level demographic characteristics were similar between the *i-Ready Instruction* schools and comparison schools selected for analysis. Schools were matched on the following variables:

- Percent students eligible for free or reduced lunch (FRL)
- Percent students limited English proficient (LEP)
- Percent students with disabilities (SWD)
- Percent students white

We selected these variables because they are known to be related to student achievement, and reliable data are available for public schools across the country, including all schools meeting the criteria for our sample.

We used a logistic regression model to compute a propensity score for each school in the *i-Ready Instruction* pool and the comparison pool (Guo & Fraser, 2010). Propensity scores range between 0 and 1 and indicate the probability that each school belonged to the *i-Ready*

Instruction group given their values on the matching variables. We used a nearest neighbor matching approach (Stuart, 2010) to match each *i-Ready Instruction* school with an available comparison school with the closest propensity score. We matched without replacement – meaning all schools could only be matched once. Following the first round of matching including all 26 *i-Ready Instruction* schools, baseline equivalence was not achieved. Baseline equivalence (BE) is the similarity in means of the matching variables between the groups as measured by a standardized effect size. An examination of the propensity scores revealed that two of the *i-Ready Instruction* schools had scores that were outside the range of values observed for both the intervention and comparison groups (i.e., the area of common support). Thus, we used trimming to remove these two schools (Stuart, 2010) and conducted the matching procedure again, achieving BE after trimming.

According to WWC guidance, school demographic variables meet baseline equivalence if they are below 0.25 for reading (WWC, 2014). Table 3 provides descriptive statistics for each matching variable for the *i-Ready Instruction* (treatment) and comparison group, as well as the effect size of the difference between the means for the final sample of study schools.

Table 3. School Demographic Variables and Effect Size Differences Between the i-Ready Instruction (Treatment) and Comparison Group.

Variable	*i-Ready Instruction* (N = 24)		Comparison (N = 24)		Adj. Mean Diff.	Effect Size
	Mean	SD	Mean	SD		
Percent FRL	55.72	23.43	51.20	19.23	4.52	0.21
Percent LEP	3.64	3.44	3.75	3.50	-0.11	-0.03
Percent SWD	17.50	4.91	16.99	3.59	0.52	0.12
Percent White	69.92	17.98	74.46	20.33	-4.54	-0.24

Note. Effect size calculated as Hedges' g with a small sample correction per WWC procedures.

Though not part of our matching criteria, we also examined the final urbanicity make-up of our *i-Ready Instruction* and comparison schools. These data reveal that a similar number of suburban schools were included in our final *i-Ready Instruction* and comparison school samples; however, there were more urban schools in the comparison set and more rural and town schools in the *i-Ready Instruction* set (see Table 4).

Table 4. Urbanicity Differences Between the i-Ready Instruction (Treatment) and Comparison Group.

Urbanicity	*i-Ready Instruction* (*n* =24)	Comparison (*n* = 24)
Rural	26.09	16.67
Suburban	39.13	41.67
Town	26.09	8.33
Urban	8.70	33.33

Student-Level Matching

Following school-level matching, we examined student-level baseline equivalence for students eligible for inclusion as treatment or comparison using only the matched schools. We observed baseline equivalence for each grade using the matched schools; therefore, student-level matching was not necessary. Table 5 provides the mean Fall 2017 reading achievement, as measured by the reading *i-Ready Diagnostic*, for the full set of students at each grade level in matched schools. The effect size differences were less than the WWC threshold of 0.25 for baseline equivalence at each grade level (WWC, 2017b). The adjusted mean difference between students at each grade level was estimated using a mixed model that nested students within schools and estimated the difference between the treatment group and the comparison group at the school level.

Table 5. Reading Baseline Equivalence Statistics for i-Ready Instruction (Treatment) and Comparison Groups by Grade

Grade	Group	Schools	Students	*i-Ready Mean*	*i-Ready SD*	Adj Mean Diff (SE)	Effect Size
6	*i-Ready Instruction*	24	3444	575.09	50.90	3.97 (4.76)	0.077
	Comparison	24	4047	568.57	51.86		
7	*i-Ready Instruction*	24	2284	585.13	52.91	-1.50 (5.33)	-0.028
	Comparison	24	3807	583.75	54.12		
8	*i-Ready Instruction*	24	1509	587.38	54.02	-11.17 (5.18)	-0.210
	Comparison	24	3718	596.57	52.88		

Note. ICC = intraclass correlation, SD = standard deviation of *i-Ready* scores, Adj Mean Diff = adjusted mean difference between *i-Ready Instruction and Comparison groups*, and SE = standard error of the adjusted mean difference.

Impact Analysis

Following the selection of baseline equivalent students, we used hierarchical linear modeling (HLM) to address our research question to estimate the impact of *i-Ready Instruction* use on student achievement in reading. We implemented a two-level model to account for the clustered nature of the data with students nested within schools. Because effect size differences between the treatment and comparison groups on student achievement at baseline fell between 0.05 and 0.25 standard deviations, we included baseline reading achievement in the model as covariates. We ran separate models at each grade level for a total of three analyses.

We specified Level 1 of the model as:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(PRE\_READ)_{ij} + e_{ij}$$

where Y_{ij} is the spring reading *i-Ready Diagnostic* score for student *i* in school *j*. β_{0j} is the adjusted mean outcome for students in school *j*. β_{1j} is the adjusted difference in outcome due to the student's baseline score in reading (school mean centered). e_{ij} is the random error in the achievement outcome associated with student *i* in school *j* not accounted for in the model.

We specified Level 2 of the model as:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{GROUP})_j + \gamma_{02}(\text{PRE\_READ}._j - \text{PRE\_READ}..)_j + \Sigma\gamma_K(\text{SCHOOL}) + u_{0j}$$

where γ_{00} is the adjusted comparison 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 *i-Ready Instruction* group and 0 for schools in the comparison group. γ_{02} is the regression slope of the school-level baseline achievement in reading (grand mean centered). γ_{02} is included to explain additional between-school variance not explained in level 1 of the model. 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 covariate used in each analysis were:

- *i-Ready Diagnostic* reading baseline performance

The school-level covariates included:

- Group membership (*i-Ready Instruction* or comparison)
- Average *i-Ready Diagnostic* baseline score in reading
- Demographic variables
  - Percent white students
  - Percent Limited English Proficient (LEP) students
  - Percent of students eligible for free and reduced-price lunch (FRL)
  - Percent of students with disabilities (SWD)

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 on outcome variables, effect sizes were computed for all comparisons using Hedge's *g* with an adjustment for small sample sizes (WWC, 2017b).

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 C.

Students in the *i-Ready Instruction* schools were not statistically significantly different (α = .05) than the students in the comparison schools on reading achievement as measured by student reading *i-Ready Diagnostic* scores for grades 7 and 8 (Table 6). However, students in *i-Ready Instruction* schools performed statistically significantly higher than matched comparison schools for grade 6.

The effect sizes, measured by Hedge's *g*, for the three grades were 0.09, 0.06, and 0.07 for grades 6, 7, and 8, respectively. All effect sizes were considered small for education interventions (Lipsey et al., 2012)

The intra-class correlations (ICCs) are also presented in Table 6. The ICCs measure the proportion of the variance that is between schools—that is, how much of the variance in reading *i-Ready Diagnostic* scores can be explained by school-level differences. The ICCs were 0.10,

0.10, and 0.08 for grades 6, 7, and 8, respectively. This suggests the majority of variance in reading achievement is due to factors other than school-level differences.

Table 6. Impact Analysis Results for i-Ready Diagnostic Schools Compared to Comparison Schools for Reading Student Achievement at grades 6–8.

Grade	Group	Schools	Students	*i-Ready Mean*	*i-Ready SD*	Adj Mean Diff (SE)	*p*-value	Effect Size	ICC
6	*i-Ready Instruction*	24	3,444	588.47	54.43	4.79 (2.32)	0.04	0.09	0.10
	Comparison	24	4,047	581.27	52.62				
7	*i-Ready Instruction*	24	2,284	598.73	53.08	3.42 (2.23)	0.13	0.06	0.10
	Comparison	24	3,807	596.61	55.51				
8	*i-Ready Instruction*	24	1,509	599.47	57.74	3.73 (2.56)	0.15	0.07	0.08
	Comparison	24	3,718	605.63	55.38				

Note. ICC = intraclass correlation; SD = standard deviation of *i-Ready* scores; Adj Mean Diff = adjusted mean difference between *i-Ready Instruction* and Comparison groups; SE = standard error of the adjusted mean difference; and Effect Size = Hedge's *g*.

Impact Summary and Discussion

HumRRO and Century Analytics' study of the impact of middle school reading *i-Ready Instruction* on student achievement found a positive significant effect at grade 6 – with students using *i-Ready Instruction* performing better than those in a matched comparison group. However, the effect size of the difference was small (0.09). At grades 7 and 8 those students using *i-Ready Instruction* performed no better than those in a matched comparison group, and effect sizes were also small.

All schools attempt to provide a valuable education to their students by implementing quality curriculum and classroom assessments. Therefore, all students should expect to see gains in student achievement. There are various possible reasons, other than ineffectiveness, why there was no statistically significant impact of *i-Ready Instruction* use on reading achievement in two of the three grades in middle school. One possibility is sampling error which is the chance that the sample does not show an effect when there is one in the population. A second possibility is that schools may have solid reading curriculum and interventions in place at the comparison schools and the *i-Ready Instruction* is comparable.

Additionally, through our review of *i-Ready* data and in conversations with Curriculum Associates staff who work directly with schools, we understand that middle schools often do not implement *i-Ready Instruction* as consistently as often seen at the elementary school-level. Though we applied rules for identifying treatment schools based on the current Curriculum Associates guidance for good implementation, the team may need to revisit and revise these rules based on the realities of middle school schedules and demands. Similarly, we understand Curriculum Associates' products have undergone recent changes to make *i-Ready Instruction* more age-appropriate to older students – particularly, Curriculum Associates has received feedback from teachers indicating the characters and games are too young for middle school students. Curriculum Associates has listened to this feedback, and developers have made updates at the middle school grade levels.

Finally, our treatment group was compared to a comparison group using the *i-Ready Diagnostic*. It is possible that use of *i-Ready Diagnostic* itself increases student achievement; however, the design of this study did not allow for an estimation of that impact. Use of the *i-Ready Diagnostic* only schools as a comparison group may have attenuated the effects of the treatment had that group been compared to a "business-as-usual" comparison group. Future studies might examine the impact of *i-Ready Instruction* use to a set of a comparison schools and students not implementing any Curriculum Associates products. This would require an external achievement measure, potentially state assessments, for use as an achievement baseline measure and outcome variable.

There are various limitations to consider when interpreting the study results. The schools and students in this study were not participants in a research study but actual customers and everyday users of educational products. We relied on implementation of *i-Ready Instruction* carried out in real-world conditions. Implementation of *i-Ready Instruction*, therefore, likely varied between schools. We may have found different results had the study been conducted under more controlled circumstances. Impacts are typically greater for studies that aim for ideal or close to ideal implementation and less for studies that examine real-world implementation.

This study was conducted using a rigorous quasi-experimental design (QED) to meet the standards described in the WWC 4.0 standards to achieve a rating of *Meets WWC Group Design Standards with Reservations.* This study was able to control for student baseline reading

achievement and school demographic characteristics. It is possible, however, that other student or school characteristics are contributing to student achievement in a way not examined by this study. We recommend a future study consider stratifying the matching of students by school such that students are matched to peers within the same school. We further recommend to only include students for whom student-level demographic data are available (e.g., disability status, English learner status, race/ethnicity) to provide an opportunity to match students using multiple variables known to be related to student achievement.

We did not find statistically significant positive results for two of the three grades. This study meets the guidelines set forth by ESSA for a *Level 2* (or *Moderate*) rating for evidence-based research at grade 6. ESSA guidance requires positive significant results to support an intervention (U.S. Department of Education, 2016).

Quality Control Procedures

We employed various quality control checks throughout the data cleaning, analysis, and reporting process. HumRRO, Curriculum Associates, and Century Analytics worked together to identify a rigorous methodology based on proper implementation of *i-Ready Instruction*, the WWC 4.0 standards, and ESSA Level 2 guidelines.

Rules for identifying treatment and comparison groups were determined upfront through collaboration between the three groups. Curriculum Associates provided information on the various components of *i-Ready Instruction*, and the frequency for which it should be used for recommended implementation. They provided *i-Ready Diagnostic* and *Instruction* data to allow HumRRO and Century Analytics to empirically examine the extent to which these recommendations were followed by *i-Ready Instruction* schools. These discussions led to a treatment and comparison school criteria that all partners were confident in.

Data analysis work was completed collaboratively by HumRRO and Century Analytics. Century Analytics and HumRRO independently conducted matching and HLM analyses for each grade. The researchers reviewed results against each other and worked out any discrepancies. All data reported in this study were verified by both researchers.

References

Curriculum Associates. (2018). *i-ReadySpring 2019: Leadership Success Guide.* Billerica, MA: Curriculum Associates, LLC. Retrieved from file:///C:/Users/rdvorak/Downloads/iready-pd-success-guide-for-leaders-2019-2.pdf

Curriculum Associates. (2019, June 20). *Getting the MOst from Online Instruction*. Retrieved from *i-Ready*Central: http://i-readycentral.com/articles/getting-the-most-from-online-instruction/

Curriculum Associates. (2019, June 27). *The Research Behind our Programs*. Retrieved from https://www.curriculumassociates.com/research-and-efficacy

Curriculum Associates. (2019a, June 27). *The Research Behind our Programs*. Retrieved from https://www.curriculumassociates.com/research-and-efficacy

Curriculum Associates. (2019b, July 22). *Get Good Data*. Retrieved from *i-Ready*Central: http://i-readycentral.com/articles/get-good-data/

Guo, S. and Fraser, M.W. (2010). Propensity Score Analysis. Thousand Oaks: SAGE.

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/.

Spybrook, S., Bloom, H., Congdon, R., Hill, C., Liu, X., Martinez, A., & Raudenbush, S. (2011) *Optimal Design Plus Empirical Evidence* (Version 3.01): W. T. Grant Foundation.

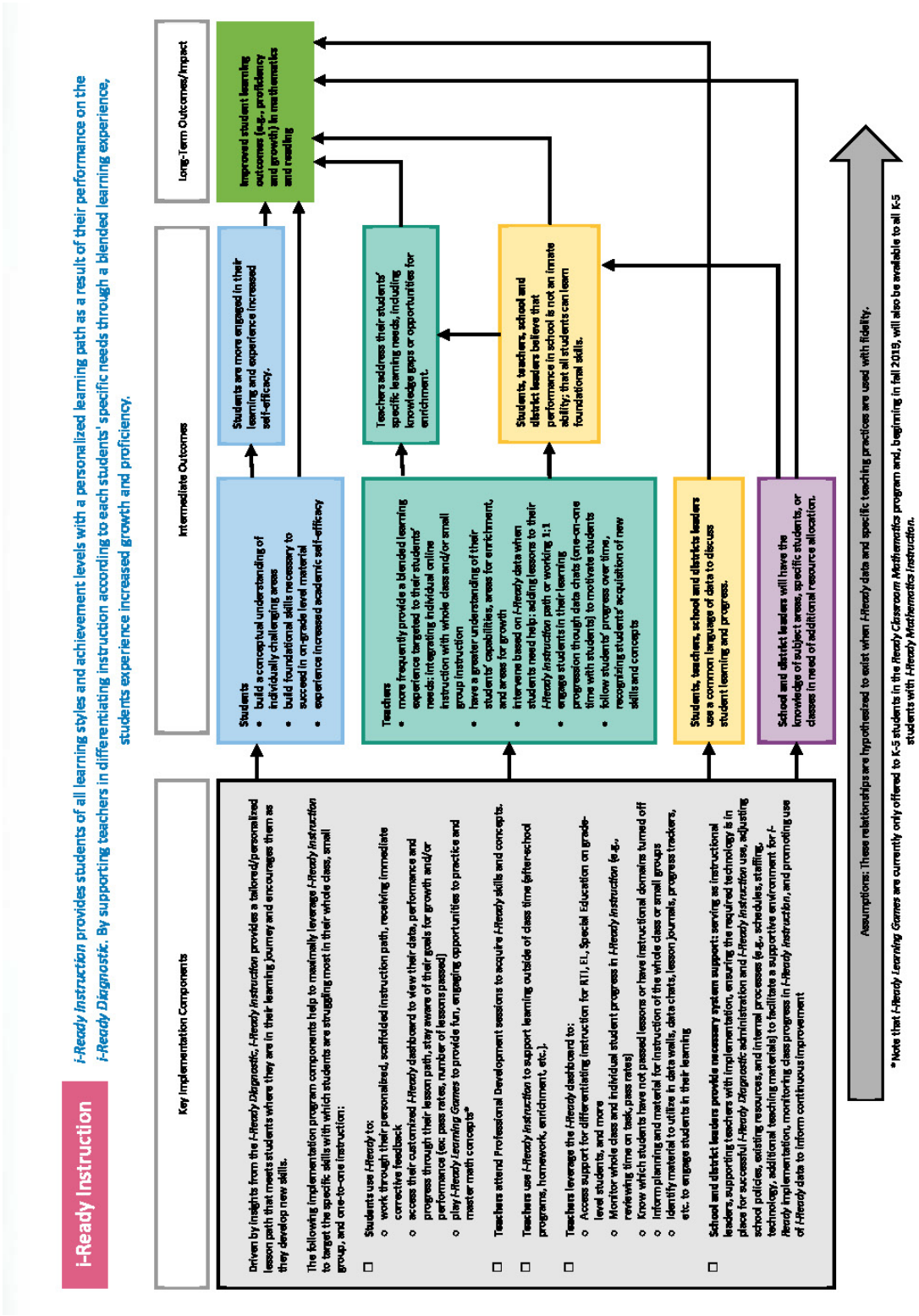
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1–21.

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. (2017a, October). *What Works Clearinghouse: Procedures Handbook (Version 4.0).* Retrieved from http://whatworks.ed.gov.

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

Appendix A. *i-Ready Instruction* Theory of Action



Appendix B.
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.

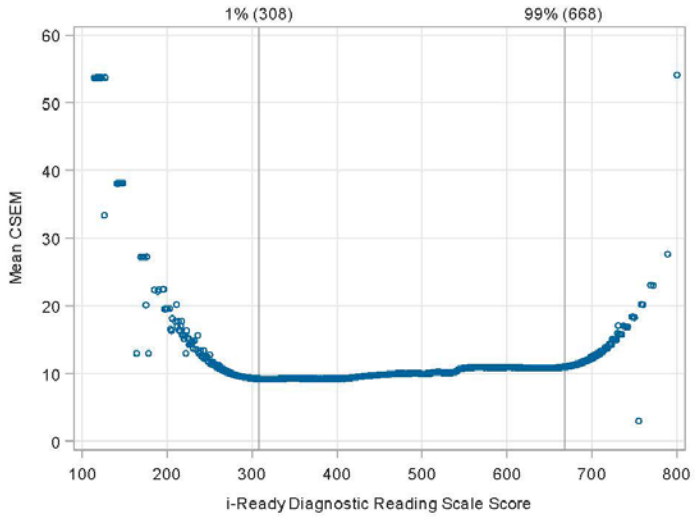


Figure 9. Conditional SEM of i-Ready Diagnostic Assessments for Reading

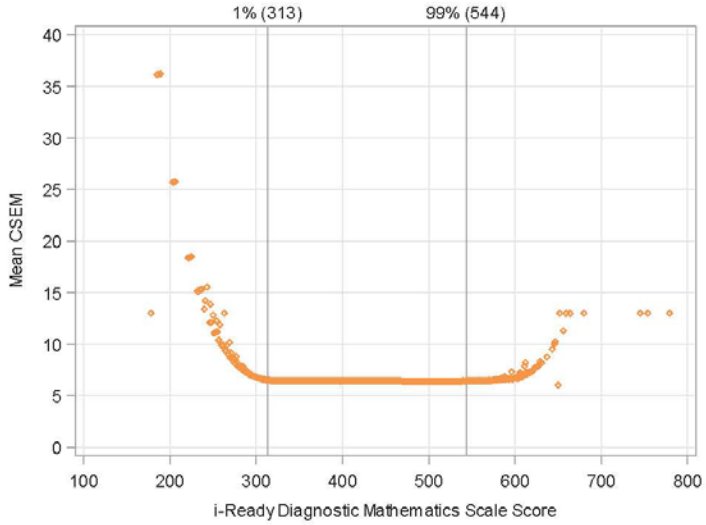


Figure 10. Conditional SEM of i-Ready Diagnostic Assessments for Mathematics

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).

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

where ρ_θ is the marginal reliability estimate, $\sigma^2_{\hat{\theta}}$ is the observed error variance of the ability estimate, and $\bar{\sigma}^2_E$ 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 C. Impact HLM Coefficients

Table C.1. HLM Results for Grade 6 Reading.

Covariates	Coef.	SE	z	*p*-value	95% Conf. Interval	
Student-Level Covariates						
Treatment Group Membership	4.79	2.32	2.07	0.039	0.25	9.33
Fall 2017 Reading *i-Ready*School Mean Centered	0.69	0.10	6.78	<.001	0.49	0.89
School-Level Covariates						
Percent LEP	131.49	62.06	2.12	0.034	9.85	253.12
Percent SWD	19.24	24.94	0.77	0.44	-29.64	68.12
Percent White	0.33	0.12	2.78	0.005	0.10	0.57
Percent FRL	-14.45	7.18	-2.01	0.044	-28.52	-0.38
Fall Reading *i-Ready*School-Level Grand Mean Centered	0.87	0.01	130.82	<.001	0.86	0.89
Intercept	557.06	12.01	46.37	<.001	533.52	580.61

Note. LEP = limited English proficient; SWD = students with disabilities; FRL = free or reduced lunch; Coef. = coefficient; SE = standard error of the coefficient; z = standardized score.

Table C.2. HLM Results for Grade 7 Reading.

Covariates	Coef.	SE	z	*p*-value	95% Conf. Interval	
Student-Level Covariates						
Treatment Group Membership	3.42	2.23	1.53	0.125	-0.95	7.80
Fall 2017 Reading *i-Ready*School Mean Centered	0.84	0.09	9.65	<.001	0.67	1.01
School-Level Covariates						
Percent LEP	114.11	64.49	1.77	0.077	-12.30	240.51
Percent SWD	-24.04	25.80	-0.93	0.351	-74.60	26.52
Percent White	0.22	0.12	1.82	0.069	-0.02	0.45
Percent FRL	-5.18	6.99	-0.74	0.459	-18.88	8.53
Fall Reading *i-Ready*School-Level Grand Mean Centered	0.85	0.01	117.21	<.001	0.84	0.87
Intercept	582.30	12.52	46.5	<.001	557.76	606.85

Note. LEP = limited English proficient; SWD = students with disabilities; FRL = free or reduced lunch; Coef. = coefficient; SE = standard error of the coefficient; z = standardized score.

Table C.3. HLM Results for Grade 8 Reading.

Covariates	Coef.	SE	z	*p*-value	95% Conf. Interval	
Student-Level Covariates						
Treatment Group Membership	3.73	2.56	1.46	0.145	-1.29	8.74
Fall 2017 Reading *i-Ready*School Mean Centered	0.78	0.09	8.85	<.001	0.61	0.95
School-Level Covariates						
Percent LEP	36.37	73.14	0.5	0.619	-106.98	179.72
Percent SWD	-43.75	28.11	-1.56	0.12	-98.84	11.33
Percent White	0.09	0.14	0.68	0.495	-0.17	0.36
Percent FRL	-2.43	6.81	-0.36	0.722	-15.78	10.92
Fall Reading *i-Ready*School-Level Grand Mean Centered	0.87	0.01	101.61	<.001	0.85	0.89
Intercept	603.37	14.21	42.48	<.001	575.53	631.21

Note. LEP = limited English proficient; SWD = students with disabilities; FRL = free or reduced lunch; Coef. = coefficient; SE = standard error of the coefficient; z = standardized score.