Examining the Impact of Amplify Reading on Student Literacy in Grades K-2

2019 Report

(December, 2019)

Stephen Newton, PhD*

Harrison Gamble*

Yu Su, PhD

Jennifer Zoski, PhD

Danielle Damico, PhD

** These authors contributed equally to this work.*

Keywords: product efficacy, remote learning, multilevel modeling, propensity weighting

Abstract

This paper presents the results of a quasi-experimental study examining the effectiveness of Amplify Reading, a digital supplemental literacy curriculum that students play independently at school or at home, in improving student literacy outcomes as measured by end of year Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Next for students in Grades K-2. Students in Amplify Reading outperformed and outgrew a comparison sample of students from their district. Significant and meaningful effects were obtained over one semester with an average of fewer than seven hours of use of the program. This is less than the expected use over the course of a school year, suggesting the potential of stronger impact when the program is used over a longer period of time.

Introduction

This study describes the impact of Amplify Reading on students' early literacy skills in kindergarten through second grades. Using data from a large urban school district, the study assessed the effectiveness of Amplify Reading K-2 by comparing the performance of consistent Amplify Reading users with a control group of similar students in similar schools who did not use Amplify Reading.

Overview of the program

This study focuses on the K-2 version of Amplify Reading. Amplify Reading is a research-based, standards-aligned digital supplemental literacy curriculum that engages and motivates students through a variety of mini-games, each focusing on building proficiency in early reading skills while providing opportunities to apply those skills in increasingly complex texts. The program was designed to include content that is most effective at building the word reading and comprehension skills of elementary students (e.g., NICHD, 2000; NIFL, 2008), including at-risk and struggling readers (e.g., NICHD, 2000) and English language learners (e.g., August & Shanahan, 2006). Because Amplify Reading is a supplemental program, the skills included allow for a balance of breadth and depth of instruction. Students in kindergarten through second grades engage in practice and explicit instruction in the underlying phonics, phonological awareness (PA), vocabulary, and comprehension skills that are essential for fluent reading with strong comprehension (e.g., Cartwright, 2010; NICHD, 2000; Oakhill, Cain, & Elbro, 2015).

Students are introduced to new skills and concepts in a variety of carefully sequenced and paced mini-games, which are delivered based on a literacy skills scope and sequence and adapt based on student performance (Refer to Table 4 in the Appendix for a list of mini-games by skill and grade). Since learning is promoted when students use knowledge across tasks (Merrill, 2002), Amplify Reading encourages generalization through ebooks with embedded activities that reinforce skills in longer, more authentic texts. Students are initially placed into the program scope and sequence using available literacy screening data. Once students are placed in the program, their movement is driven by their performance within the Amplify Reading mini-games and ebooks. In this way, the program adapts to meet students where they are, providing them with instruction and practice in skills that they are ready for, while challenging them with increasingly more difficult content as they progress.

The skills that make up the Amplify Reading K-2 literacy map can be broken down into three main categories: phonological awareness and phonics, comprehension, and vocabulary. The research behind the specific targeted skills in these areas is described next in more detail.

PA and Phonics

Much research has documented the importance of foundational skills (the skills required to decode words) and the impact of instruction in foundational skills on overall reading success and long-term student outcomes. Students' mastery of the code is causally related to comprehension (e.g., Garcia & Cain, 2014; McCandliss, Beck, Sandak, & Perfetti, 2003). Decoding skill is essential for reading new words and developing reading fluency, but the opacity of English makes it one of the most difficult orthographies to learn (Aro & Wimmer, 2003; Ellis *et al*., 2004; Wimmer & Goswami, 1994). Phonics instructional approaches help students crack the code by highlighting spelling regularities and giving students rules for letter-sound correspondences so that they are able to decode new words, building toward automatic word recognition.

In order to convert written words into speech or decode, students must master many skills. Most broadly, those skills fall into the domains of phonological awareness and phonics/word analysis. Phonological awareness refers to an understanding that words are made up of sounds and the ability to manipulate the sounds in words, from syllables (e.g., tell me the parts of *jacket, jack-et*) to phonemes, the smallest unit of sound (e.g., tell me the sounds in *cat*, /c/ /a/ /t/*).* Phonological awareness skills are necessary for students to decode text (Smith, Simmons, & Kame'enui, 1998; Torgesen, Wagner, & Rashotte, 1994). Students must also master skills in the phonics domain, from the sounds that individual letters and letter combinations make (e.g., *b* says /b/, *oo* says /oo/) to the use of strategies for breaking words into parts (roots, prefixes, suffixes, syllables) to read them (e.g., look for the root and ending in *jumped* to read the word). Further, it isn't enough for students to demonstrate accuracy with these skills; they must also be able to engage in phonological awareness and decoding skills with a level of fluency or automaticity to facilitate fluent reading for meaning (Hudson, Pullen, Lane, & Torgesen, 2009; Ritchey & Speece, 2006).

Amplify Reading provides students with explicit instruction and practice across the continuum of phonological awareness and phonics skills that students must learn to become successful decoders. The program covers the foundational skills included in the Common Core State Standards (National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010), giving extra emphasis to those that have been consistently documented as predictive of future reading success.

Comprehension

The comprehension instruction in Amplify Reading is grounded in the most current research on what strong readers do to make meaning from text. Comprehension instruction often focuses on the products of good comprehension (demonstrations of understanding after reading is complete) rather than the processes of comprehension (the activities a reader does to comprehend text during reading) (Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007). However, a large body of research has documented the underlying skills that are critical for reading comprehension (e.g., Cartwright, 2010; Oakhill et al., 2015). These are the skills that are necessary for building a mental model or a network of idea units that readers construct in order to comprehend the gist of what they are reading (e.g., Graesser, Singer, & Trabasso, 1994; Kintsch, 1988). Students who struggle with reading comprehension are often weak in the underlying language and literacy abilities that are required to create this coherent mental model (e.g., Cartwright, 2010; Oakhill et al., 2015); these underlying skills are collectively referred to as microcomprehension.

Leading researchers Oakhill and Cain catalogued these model-building skills using the term 'inference' (e.g., Oakhill et al., 2015). At roughly the same time, Graesser at the University of Memphis was exploring the same topic from the direction of ‘coherence:’ good texts have coherence (they aren't just collections of unrelated sentences) but poor mental models lack it (Graesser, McNamara, & Louwerse, 2002). For the purposes of understanding the full research base to develop programs that effectively teach the skills students need to successfully build mental models, the work of these leaders in the field and others has been combined under the umbrella term ‘microcomprehension.’ These comprehension skills include: cognitive flexibility, syntactic awareness, connectives, anaphora resolution, inferencing, understanding of text structures, and awareness of text schema.

In Amplify Reading, microcomprehension instruction is provided in addition to instruction that includes work on macrocomprehension skills. These skills are addressed in multiple contexts across mini-games and ebook tasks within Amplify Reading. When explicit instruction is required, students engage in mini-games that include models of the skill with think-aloud type instruction and clear and consistent feedback. These mini-games give students opportunity to practice these critical skills with increasingly challenging texts with the goal of helping them gain and practice the skills needed to build coherent mental models of text that promote strong comprehension.

Vocabulary

Although it is not possible to teach students the meanings of all the words that they will encounter, it is critical to directly teach high-utility vocabulary that students will need to understand in order to access the curriculum. Amplify Reading provides practice with Tier 2 vocabulary because these words are often encountered during reading, are likely to be related to concepts that children will understand, and are not typically learned as part of conversational language (McKeown, Beck, & Sandora, 2012).

In addition to high-utility Tier 2 words, students receive instruction and practice with multiple-meaning words. Readers who can think flexibly about word and sentence meanings are better able to use context to monitor their comprehension for meaning (Zipke, Ehri, & Smith Cairns, 2009). Students in Amplify Reading build this skill through riddles. They practice choosing correct punchlines for jokes and determining why ambiguous words make jokes funny, an exercise that has been shown to improve students’ reading comprehension (Yuill, 2009).

Vocabulary instruction in Amplify Reading encourages depth of knowledge by having students explore the relationships among words across multiple contexts (Beck, Perfetti, & McKeown, 1982; Coyne, McCoach, & Kapp, 2007; McKeown, Beck, & Sandora, 2012). Students build this knowledge by organizing related vocabulary by semantic gradients or shades of meaning. By arranging vocabulary words in an order that represents subtle differences in meaning, students are able to connect known vocabulary words with new ones and see the relationships graphically (Blachowicz & Fisher, 2015). The goal is for students to build the depth of their vocabulary knowledge by strengthening connections between words, allowing them to build stronger and larger vocabulary networks.

Students also receive morphological analysis instruction and practice to build a word-learning strategy that will help them acquire the meanings of new words independently during reading (Bowers, Kirby, & Deacon, 2010; Goodwin & Ahn, 2010). Focusing morphological instruction on frequently occurring, consistently spelled morphemes gives students a powerful strategy to use when tackling unknown multimorphemic words during reading, and is thus an efficient way to improve students' vocabulary and reading comprehension skills (e.g., Bowers *et al*., 2010; Goodwin & Ahn, 2010).

Study Design

The study addressed the following research question:

> What is the impact of Amplify Reading on the reading skills of students in Kindergarten through grade 2?

In other words, this is a study of Amplify Reading's effectiveness in improving student literacy. In making such a causal claim, in the potential outcomes framework, we want to know the difference in student outcomes for students who used the product versus the counterfactual condition, that is, what those same students would have scored had they not used the product. By definition, counterfactuals cannot be directly observed because they did not occur. So, researchers look for other participants to function as stand-ins when measuring the counterfactual result. Randomized experiments provide the best way to estimate the counterfactual, since random assignment creates treatment and comparison groups that are, on average, the same on measured and unmeasured characteristics prior to receiving an intervention.

Not having the opportunity for random assignment, this study used two steps to compare Amplify Reading students with similar students from similar schools in the comparison group. First, we used propensity score weighting to identify a comparison group. The propensity score is defined as a conditional probability of treatment assignment, given observed explanatory variables (Rosenbaum & Rubin, 1983). This weighting assured that inferences about causation were based on similar types of students in the treatment and comparison group. Without such weighting, treatment and comparison groups are typically quite different and can only be compared with a reliance on model-based statistical controls, which can result in biased estimates whenever extrapolation is inaccurate. With propensity score weighting, however, the comparison group closely resembles the treatment group across all measured variables, and so can be used as a reasonable proxy for the treatment group. Second, our outcome model controlled for student-level and school-level explanatory variables to remove bias from any remaining measured differences between Amplify Reading students and comparison students.

Sample

The study took place in a large urban school district in the 2018-19 academic year. All students in Kindergarten, 1st and 2nd grades at Title 1 schools were given the option to use Amplify Reading. The study population selected in each grade was highly diverse along a number of different dimensions; 80% of students were Hispanic, 33% were English Learners and 10% received special education services (For a complete table of student demographics, see Table 1 in the Appendix). We carried out the analysis outlined below separately on students in each grade level, because the measure we used to measure literacy ability, DIBELS Next, is not vertically scaled.

Treatment group

To answer our research question, we estimated the average effect of a treatment on the treated (ATT). In this approach, it is necessary to identify a treatment group of students who used the product as intended. Since most schools in this district began using Amplify Reading in the spring semester (i.e., early 2019), we focused on students who used the product consistently between their middle-of-year (MOY, starting in about January) and end-of-year (EOY, ending in June) DIBELS assessments. Between these tests, students typically received 16 weeks of instruction. Given that the recommended usage of Amplify Reading is 30 minutes per week, students who met their targets every week would have used the product 8 or more hours. Our research question did not focus solely on the effect of *ideal* implementation of Amplify Reading. To make our results more generalizable, we included students in the treatment group who did not meet the optimal desired usage but still used the product regularly. We defined the treatment group as all students who used Amplify Reading for 4 or more hours, which averages to 15 minutes or more per week between a student's MOY and EOY assessments. 16,669 students were considered treated using this definition, with 4,647 students, 6,041 students, and 5,981 students in Kindergarten, Grade 1, and Grade 2 respectively. The treatment group used Amplify Reading for an average of 7 hours and 44 minutes, meaning the average student in the treatment group was close to the target of 8 hours, though some users fell below the target and others surpassed it.

Comparison Group

Our comparison group was defined using all students who did not use Amplify Reading between MOY and EOY. This choice of comparison, in conjunction with the preceding definition of treatment, excluded all students who used Amplify Reading but fell short of 4 hours total usage. To establish group equivalence, we first considered using propensity score matching, but after identifying a matched sample, we found that the treatment and comparison groups still differed by more than the conventional criterion of .10 standard score units on many variables. When standardized differences exceed .10, proportional odds weighting is a more appropriate approach for creating similar treatment and comparison groups in an ATT analysis (Leite, 2017), so we used that approach for defining the comparison group instead. Furthermore, propensity score weighting has been found to produce more precise estimates of treatment effects and remove more selection bias compared to propensity score matching (Leite, Aydin & Gurel, 2019), particularly when balancing across many explanatory variables (Elze *et al*., 2017).

Measures

Students in both treatment and control groups were administered the mCLASS: Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Next assessment at the MOY and EOY during the 2018-2019 school year. DIBELS Next is a nationally recognized screening assessment developed by the University of Oregon for assessing the acquisition of early literacy skills from kindergarten through grade 6. In kindergarten through second grade, the assessment is administered one-on-one with students by a qualified professional and includes measures that serve as indicators of reading skills such as alphabet knowledge, phonemic awareness, alphabetic principle/phonics, fluency, and reading comprehension. Skills assessed vary by grade level and time of year. DIBELS Next has strong reliability for individual measures and the overall composite score (Good *et al*., 2013).

The DIBELS Next measures that were administered at each benchmark period were specific to the students' grade and time of year, progressing from measures of lower-level phonological awareness and phonics to measures of higher-level fluency and comprehension skills. The Composite Score for each testing session is a combination of multiple DIBELS measures, which vary by grade and benchmarking period (see Table 1 below for a breakdown). Overall, the composite score provides the best overall estimate of the student's reading proficiency and risk level (Good *et al*., 2013). Evidence for the reliability and validity of each of the measures that contributed to the Composite Score, as well as for the Composite Score itself is presented in the following sections.

Table 1: Measures Contributing to the Composite Score at Middle of Year (MOY) & End of Year (EOY)

KINDERGARTEN	
MOY	***EOY***
FSF	
LNF	LNF
PSF	PSF
NWF-CLS	NWF-CLS
FIRST GRADE	

MOY	***EOY***
NWF-CLS	
NWF-WWR	NWF-WWR
DORF Words Correct	DORF Words Correct
DORF Accuracy	DORF Accuracy
SECOND GRADE	
MOY	***EOY***
DORF Words Correct	DORF Words Correct
DORF Accuracy	DORF Accuracy
Retell	Retell

First Sound Fluency (FSF)

FSF is a test of a student's fluency in identifying the initial sounds in words. It is used as a measure of developing phonemic awareness at the beginning and middle of kindergarten. Students are asked to say the first sound in a word for as many words as possible in one minute. Among kindergarten students, the authors report a 2-week single-form alternative reliability of 0.85, a 2-week three-form alternate form reliability of 0.95, and inter-rater reliability of 0.95, indicating strong reliability for the measure (Dewey, Powell-Smith, Good, & Kaminski, 2015). Predictive validity, as measured by the correlation between FSF and the Group Reading Assessment and Diagnostic Evaluation (GRADE) end of year test was 0.52; the correlations with the Comprehensive Test of Phonological Processing (CTOPP) at the end of year ranged from 0.19 to 0.49, suggesting moderate to strong validity (Good *et al*., 2013).

Letter Naming Fluency (LNF)

LNF is a brief, direct measure of a student's automaticity with letter naming. Students are asked to identify and name uppercase and lowercase letters arranged in a random order. The total score is the number of correct letter names that the student says in 1 minute. The authors report alternate form reliability from 0.86 to 0.95 and inter-rater reliability from 0.99 to 1.00, suggesting strong reliability. (Good *et al*., 2013). Predictive validity, as measured by the correlation with the GRADE end of year assessment, ranged from 0.35 to 0.39, suggesting moderate validity.

Phoneme Segmentation Fluency (PSF)

PSF is a phonological awareness measure that assesses a student's fluency in segmenting spoken words into their component phonemes or sound segments. After hearing a word read aloud, students must verbally produce the individual sounds for each word. The total score is the number of correct sound segments that the student says in 1 minute. Students are given partial credit for partial segmentation. For example, a student who segments the word *sun* into /s/ /un/ receives 2 points, and a student who segments it into /s/ /u/ /n/ receives the full 3 points for the word. The authors report PSF alternate form reliability from 0.44 to 0.78 and inter-rater reliability from 0.95 to 0.98, suggesting strong reliability. Predictive validity, as measured by the correlation with the GRADE end of year assessment ranged from 0.24 to 0.34, suggesting moderate validity evidence (Good *et al*., 2013).

Nonsense Word Fluency (NWF)

NWF is a brief, direct measure of the alphabetic principle and basic phonics. It assesses knowledge of basic letter-sound correspondences and the ability to blend letter sounds into consonant-vowel-consonant (CVC) and vowel-consonant (VC) words. The test items used for NWF are phonetically regular nonsense words. To successfully complete the task, students must rely on their knowledge of letter-sound correspondences to blend sounds into whole nonsense words. There are two separate scores reported for NWF: Correct Letter Sounds (CLS) and Whole Words Read (WWR). CLS is the number of letter sounds produced correctly in one minute. WWR is the number of nonsense words read correctly as a whole word in one minute. To complete this measure, students are presented with a sheet of randomly ordered VC and CVC nonsense words (e.g., dif, ik, nop) and are asked to read them as best as they can, reading either the whole word or saying the sounds they know. The authors report NWF-CLS alternate form reliability from 0.71 to 0.94, inter-rater reliability from 0.99 to 1.00, and test-retest reliability from 0.76 to 0.90 (Good *et al*., 2013). Predictive validity, as measured by the correlation with EOY GRADE assessment ranged from 0.43 to 0.56. For NWF-WWR, alternate form reliability ranged from 0.90 to 0.97, inter-rater reliability ranged from 0.99 to 1.00, and test-retest reliability ranged from 0.70 to 0.88. Predictive validity for NWF-WWR, as measured by the correlation with the EOY GRADE assessment, ranged from 0.39 to 0.56. Together, this suggests that NWF-CLS and NWF-WWR have moderate to strong validity and strong reliability evidence.

DIBELS Oral Reading Fluency (DORF)

DORF is a test of accuracy and fluency with connected text. Students are given an unfamiliar, grade-level passage of text and are asked to read aloud for 1 minute. For each benchmark assessment, students are asked to read three different grade-level passages for 1 minute each. Two student scores are calculated: number of words correct and accuracy rate. The number of words correct is the median number of words read correctly (with no errors, such as substitutions, omissions, or hesitations for more than 3 seconds) across the three passages. The student's accuracy is calculated by dividing the median words read correctly by the sum of the median words read correctly and the median number of errors. The authors report alternate form reliability from 0.88 to 0.98, inter-rater reliability from 0.91 to 0.95, and test-retest reliability of 0.99, suggesting strong validity (Good *et al*., 2013). Predictive validity, as measured by the correlation with the EOY GRADE assessment ranged from 0.59 to 0.77; the correlations with NAEP Oral Reading Passage ranged from 0.83 to 0.97, suggesting strong validity evidence.

DORF Retell

A passage retell component follows the reading of each DORF passage, provided that the student has read at least 40 words correct per minute on that passage, or if the assessor feels it is otherwise appropriate. Passage retell is intended to provide a comprehension check for the DORF assessment, providing an indication that the student is reading for meaning. During Retell, the student is asked to tell about what he/she has read and the assessor keeps track of the number of words in the Retell that are related to the story. The assessor also makes a judgement about the quality of the response based on how well the student retold the passage to get a qualitative rating of the student's response. The authors report that for second grade students, alternate form reliability was 0.68, test-retest reliability was 0.27, and inter-rater reliability was 0.98, suggesting moderate to strong reliability for this measure (Good et al., 2013). Predictive validity for second grade students, as measured by the correlation with the EOY GRADE assessment was 0.48, suggesting moderate validity evidence.

DIBELS Composite Score

For the Composite Score, the authors report alternate form reliability from 0.66 to 0.97, inter-rater reliability from 0.81 to 0.94, and test-retest reliability from 0.97 to 0.99, suggesting strong reliability evidence (Good *et al*., 2013). Predictive validity, as measured by the correlation with the Group Reading Assessment and Diagnostic Evaluation (GRADE) end of year test ranged from 0.50 to 0.80, suggesting strong validity evidence.

Data Analysis

Propensity Score Weighting

Each student in the comparison group was assigned a propensity score using the fitted value from a multilevel logistic model whose definition can be found in definition 1, expression 2, and equation 3 below:

$$_{ij} \equiv \Pr(y_{ij} \,|_{ij}, w_j, \mu_j) \quad (1)$$

$$y_{ij} \,|_{ij} \sim \text{Binomial}(1, \; _{ij}) \quad (2)$$

$$logit(_{ij}) \; = \; + \; \beta_{ij} \; + \; w_j + \mu_j \; (3)$$

We used a two-level hierarchical generalized linear model (HGLM, which in this case is a two-level model for binary outcomes, Raudenbush & Bryk, 2002) to estimate the propensity of being in the treatment group for student *i* in school *j* based on various student and school level characteristics variables. Specifically, definition (1) defines the probability of student *i* in school *j* being in the treatment conditioning on student and school characteristics along with school random effects as $_{ij}$ = $\Pr(y_{ij} \,|_{ij}, w_j, \mu_j)$. Expression (2) specifies that the observed treatment status (i.e., y) takes on a value of 1 given $_{ij}$ follows a binomial distribution.

Equation 3 is a random intercept logistic regression model estimating student-level likelihood of being in the treated group.   represents the average likelihood of receiving treatment when all predictors are equal to 0 (i.e., cases in the reference group for all categorical predictors also scoring at the mean for continuous predictors). μ_j represents a school-level random effect. **β** represents a vector of student-level fixed effects for the following predictors: MOY DIBELS Composite Score, Gender, Race, Special Ed, EL status, academic days between tests, and X_{ij} represents characteristics of student (*i*) nested within school (*j*). γ represents a vector of school-level fixed effects for the following predictors: Charter, Magnet, # of students, and Teacher-student ratio, and, w_j represents school-level characteristics for school *j*.

After estimating student-level propensity scores, we then transformed each student's score into a case weight using weighting by odds. When using weighting by the odds to estimate the ATT, treated students receive a weight of one, and untreated students receive a weight equal to their odds of treatment**.** The formula for transforming propensity scores is expressed in equation 4 below.

$$W_i = \frac{P(Treatment)_i}{(1-P(Treatment)_i)} \; ; \; W_i \in [0, \infty) \quad (4)$$

As we see from equation 4, a student's weight increases as their probability of treatment increases. For example, a student with a probability of treatment of .75 would be assigned a weight of 3 (*i.e.*, .75/.25), while a student with a probability of treatment of .8 would be assigned a weight of 4 (*i.e.*, .80/.20). Transforming the propensity scores into odds of treatment means our case weights have no upper bound. As a result, some students may receive extremely large weights. We chose to truncate extreme weights to the 99th percentile of the full weight distribution as this has been found to reduce the standard errors of the estimates (Lee, *et al.*, 2011).

The two-step procedure for establishing group equivalence outlined above has been shown to control for selection effects when there is a hierarchical treatment assignment mechanism (Leite, 2017). It does so by assigning high odds of treatment to comparison students who resemble the treated group on student characteristics and who were taught in similar (or the same) schools. That is, propensity score weights adjust the distributions of explanatory variables so that they are similar across treated and untreated groups.

Outcome Modeling

In order to estimate the effect of consistently using Amplify Reading, we compared the performance of the treated and comparison groups, using a weighted, random intercept linear regression. The propensity weights calculated from equation 4 in the preceding section were incorporated into the model fitting procedure. To make our estimates of the ATT of Amplify Reading doubly robust, we chose to control for the same student and school-level characteristics we used to create the propensity scores. To estimate the treatment effect, we added a vector of indicators $I(T = 1)$ to represent whether a student was in the treatment or comparison group. The model definition is listed in equation 5 and expression 6 below.

$$EOY\ \ Composite\ Score_{ij} \ = \ + {}_{T\,ij} \ + \ \beta_{ij} \ + \ w_j + \mu_j + e_{ij} \ (5)$$

$$e_{ij} \sim Normal(0,\ \sigma^2{}_{ij});\ \mu_j \sim Normal(0,\ \sigma^2{}_{j}) \ (6)$$

Similar to the model described in equation 3,   represents the intercept, ${}_T$ represents the additional achievement associated with the indicator of treatment status for student *i* in school *j*. μ_j represents a school-level random effect. **β** represents a vector of student-level fixed effects for the following predictors: MOY DIBELS Composite Score, Gender, Race, Special Ed, EL status, academic days between tests, and X_{ij} represents characteristics of student (*i*) nested within school (*j*). γ represents a vector of school-level fixed coefficients for the following predictors: Charter, Magnet, # of students, and Teacher-student ratio, and, w_j represents school-level characteristic for school *j*, and e_{ij} represents the error term.

Results

We used the Propensity Score Weighting approach described in the above section to create an equivalent control group from our study population. We balanced the analytic sample across a number of demographic characteristics as well as key instructional variables: DIBELS Composite Score prior to treatment, the number of instructional days between DIBELS Assessments, and Student to Teacher Ratio. The mean of these variables in the Treatment and Comparison groups, in addition to the Standardized Difference the groups means are displayed in Table 2 below.

Table 2: Explanatory Variable Balance Table after Propensity Weighting

	Kindergarten			**Grade 1**			**Grade 2**		
Variable	Tx Mean	Comparison Mean	Standardized Difference in Means	Tx Mean	Comparison Mean	Standardized Difference in Means	Tx Mean	Comparison Mean	Standardized Difference in Means
Charter	0.02	0.01	0.04	0.01	0.01	0.06	0.02	0.01	0.04
Days Between Assess	0.27	0.11	0.17	0.22	0.12	0.11	0.13	0.06	0.07
Female	0.48	0.47	0.03	0.48	0.48	0.01	0.49	0.46	0.06
Is Esl	0.41	0.42	0.02	0.34	0.34	0.01	0.29	0.35	0.13
Is Special Ed	0.1	0.13	0.12	0.09	0.14	0.2	0.08	0.12	0.18
Magnet	0.1	0.1	0.02	0.12	0.13	0.02	0.14	0.16	0.07
MOY Score	0.16	0.05	0.1	0.01	-0.08	0.09	0.1	-0.09	0.2
African American	0.06	0.06	0.02	0.06	0.08	0.07	0.07	0.07	0.03
Hispanic	0.85	0.87	0.05	0.85	0.84	0.05	0.84	0.84	0.01
Other	0.05	0.04	0.03	0.04	0.04	0.02	0.04	0.04	0.01
White	0.04	0.03	0.04	0.04	0.05	0.02	0.05	0.05	0
Student Teacher Ratio	-0.11	-0.03	0.08	-0.08	-0.08	0	-0.07	-0.05	0.02
Students	-0.03	0.06	0.08	-0.12	-0.03	0.08	-0.06	0.09	0.16

Table 2 shows that the difference in means across all measured explanatory variables in all grades is at or below .2 standard deviations; differences of less than .25 standard deviations have been shown to yield unbiased estimates following regression adjustments in a subsequent outcome model (Stuart, 2010; Rubin, 2001; US Department of Education, n.d.). Furthermore, the majority of the difference in explanatory variable means were below .1 standard deviations, a threshold which indicates no significant existence of bias between groups (Austin, 2011). These findings demonstrate that our treatment and control samples are sufficiently similar across both demographic and instructional variables to proceed with our analysis of student outcomes.

Multilevel Model Predicting Student End-of-Year DIBELS Composite Scores

The multilevel model shown in Table 3 below shows the student and school level effects of Amplify Reading Treatment on students End-of-Year (EOY) DIBELS Composite Score. This model controls for the same demographic and instructional variables which we balanced in the preceding propensity analysis. Across all three grades, we see that students who received the Amplify Reading Treatment outperformed their peers with effect sizes[1] of .11, .06, and .11 respectively. We standardized students EOY Composite Scores prior to fitting the model, meaning the Amplify Reading Tx Coefficient is equivalent to the estimated effect size of treatment. For a full list of student MOY and EOY scores, refer to Tables 2 and 3 in the Appendix.

Table 3: Multilevel Model Coefficients (Standardized)

	Kindergarten			Grade 1			Grade 2		
	Estimate	**Std Error**	**T-Statistic**	**Estimate**	**Std Error**	**T-Statistic**	**Estimate**	**Std Error**	**T-Statistic**
(Student Level)									
Amplify Reading Tx	0.11	0.01	15.03***	0.06	0.01	11.09***	0.11	0.01	20.78***
Moy Score	0.82	0.00	237.21***	0.98	0.00	330.94***	0.90	0.00	307.76***
Moy Score Squared	NA	NA	NA	-0.19	0.00	-83.14	NA	NA	NA
Female	-0.02	0.01	-2.97	0.01	0.00	2.71**	0.01	0.00	2.25*
African American	0.02	0.01	1.31	-0.03	0.01	-3.19	-0.02	0.01	-2.08
Other	0.14	0.02	8.94***	0.07	0.01	5.37***	0.01	0.01	0.79
White	0.00	0.02	0.08	0.05	0.01	3.91***	-0.02	0.01	-2.06
English Learner	-0.03	0.01	-4.45	0.01	0.01	2.37**	-0.01	0.01	-1.40
Special Ed	-0.20	0.01	-18.46	-0.08	0.01	-9.60	-0.17	0.01	-19.9

[1] These effect sizes are the regression coefficients for the Amplify Reading Treatment variable after estimating our multilevel outcome models. Given that EOY Composite Scores were standardized prior to the regression analysis, these effect sizes are equivalent to Cohen's d.

Days Between Assess	0.04	0.01	7.32***	0.03	0.00	8.26***	0.04	0.00	12.31***
(Intercept)	0.02	0.02	1.04	0.16	0.01	15.8***	-0.05	0.01	-5.95
(School Level)									
Charter	-0.02	0.09	-0.26	-0.05	0.06	-0.85	0.06	0.05	1.22
Magnet	0.00	0.04	0.07	0.00	0.02	0.09	0.01	0.02	0.46
Student/Teacher Ratio	0.00	0.01	0.25	0.00	0.01	-0.64	0.00	0.01	0.61
# of Students	0.00	0.01	0.29	-0.01	0.01	-1.20	-0.01	0.01	-0.79
Sd Of School Intercepts	0.18	NA	NA	0.11	NA	NA	0.10	NA	NA

*** = p < .05, ** = p < .01, *** = p < .001**

Discussion

This study was designed to identify the impact using Amplify Reading had on students during the 2018-2019 school year. To answer this question, we first identified a level of usage that we considered sufficient based on the guidance Amplify provides to teachers and administrators. We then successfully constructed a suitable control group for the students who cleared that usage threshold by weighting the population of students who did not use Amplify Reading according to their individual- and school-level similarity to our treated population. The success of this weighting procedure then allowed us to calculate an unbiased estimate of the Average Treatment Effect on the Treated (ATT). Using a multilevel regression model, we found that Amplify Reading had a significant positive effect on students End-of-Year DIBELS Composite Scores in Kindergarten, 1st and 2nd grade with effect sizes of .11, .06 and .11 respectively.

There are good reasons to believe that these effects understate the actual effects of the program, and future research will be conducted to test this notion. First, this study examined the effects of a half-year's usage (MOY to EOY). If the program is just as effective for the first part of the year, we would expect to see up to twice the effect on growth when we study change between BOY and EOY. Second, we found that MOY and EOY scores on DIBELS Composite were very closely related.[2] In this situation, typical measures of effect size tend to downplay the effects of an intervention. A more policy-relevant alternative is to estimate the effect of an intervention relative to average growth rather than relative to overall achievement (Soland and Thum, 2019), and effect sizes tend to be larger when estimated in this way. Third, product usage may have been lower because we studied the first year of implementation, and we anticipate more consistent usage as teachers and schools can build on their experience on how best to use the product.

[2] The strong relationship between MOY and EOY DIBELS Composite is indicated by the Beta coefficients from our prediction model, which were .82, .87, and .90 for grades K-2, respectively. Note the grade 1 coefficient differs from the results presented in this study because we removed the quadratic term.

In the 2019-2020 school year, we plan to explore the impact of a full year of program use in grades K-5 by leveraging an expanded sample. We also plan on identifying Amplify Reading's effect on specific skill families through a wider range of literacy outcome measures. Our planned next steps will enable us to both broaden and deepen the evidence that Amplify Reading is an effective literacy program for diverse sets of students.

References

August, D., & Shanahan, T. (2006). *Developing literacy in second language learners: Report of the National Literacy Panel on language-minority children and youth.* Mahwah, NJ: Lawrence Erlbaum Associates.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate Behavioral Research, 46*(3), 399-424.

Beck, I.L., Perfetti, C.A., McKeown, M.G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74*(4), 506-521.

Blachowicz, C. & Fisher, P.J. (2015). Learning vocabulary in the content areas. In *Teaching Vocabulary in All Classrooms*. Boston: Pearson.

Bowers, P., Kirby, J., Deacon, S.H. (2010). The effects of morphological instruction on literacy skills: A systematic review of the literature. *Review of Educational Research, 80*, 144-179.

Cartwright, K.B. (2010). *Word Callers: Small-Group and One-to-One Interventions for Children Who Read" but Don't Comprehend*. Portsmouth, NH: Heinemann.

Coyne, M.D., McCoach, B., & Kapp, S. (2007). Vocabulary intervention for kindergarten students: Comparing extended instruction to embedded instruction and incidental exposure. *Learning Disabilities Quarterly, 30*, 74-88.

Dewey, E. N., Powell-Smith, K. A., Good, R. H., & Kaminski, R. A. (2015). DIBELS Next Technical Adequacy Brief. Eugene, OR: Dynamic Measurement Group, Inc.

Ellis, N., Natsume, M., Stavropoulou, K., Hoxhallari, L., van Daal, V., Polyzoe, N., Tsipa, M, & Petalas, M. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly, 39*, 438–468.

Elze, M.C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G.W., & Pocock, S. (2017). Comparison of propensity score methods and covariate adjustment evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology, 69*(3), 345-357.

García, R. & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research.* 84. 74-111.

Good, R. H., Kaminski, R., Dewey, E., Walin, J., Powell-Smith, K., & Latimer, R. (2013). *DIBELS Next Technical Manual*, Eugene, OR: Dynamic Measurement Group, Inc.

Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. Annals of Dyslexia, 60, 183–208.

Graesser, A.C., McNamara, D.S., & Louwerse, M.M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository texts? In A.P. Sweet & C.E. Snow (Eds.), *Rethinking reading comprehension.* (pp. 82-98). New York: Guilford Press.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371-395.

Hudson, R.F., Pullen, P.C., Lane, H.B., & Torgesen, J.K., (2009). The complex nature of reading fluency: A multidimensional view. *Reading & Writing Quarterly, 25*(1), 4-32.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(4), 163-182.

Lee B.K., Lessler J., Stuart E.A. (2011). Weight trimming and propensity score weighting. *PLoS ONE, 6*(3): e18174. https://doi.org/10.1371/journal.pone.0018174.

Leite, W. (2017). *Practical propensity score methods using R.* Thousand Oaks, CA: Sage Publications.

Leite, W., Aydin, B., & Gurel, S. (2019). A comparison of propensity score weighting methods for evaluating the effects of programs with multiple versions. *The Journal of Experimental Education, 87*(1), 75-88.

McCandliss, B.D., Beck, I., Sandak, R., & Perfetti, C. (2003). Focusing attention on decoding for children with poor reading skills: A study of the Word Building intervention. Scientific Studies of Reading.7(1),75-105.

McKeown, M.G., Beck, I.L., & Sandora, S. (2012). Direct and rich vocabulary instruction needs to start early. In E.J. Kame'enui & J.F. Baumann (Eds.), *Vocabulary Instruction: Research to Practice.* New York: The Guilford Press.

Merrill, M.D. (2002). First principles of instruction. *Educational Technology, Research, and Development, 50*(2), 43-59.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards for English Language Arts*. Washington DC: Author.

National Institute for Literacy (NIFL). (2008). *Developing early literacy: Report of the National Early Literacy Panel. Executive Summary*. Washington DC: U.S. Government Printing Office.

National Institute of Child Health and Human Development (NICHD).) (2000). *Report of the National Reading Panel. Teaching Children to read: An evidence-based assessment of scientific research literature on reading and its implications for reading instruction. Report of the subgroups* (NIH Publication No. 00-4754). Washington DC: US. Government Printing Office.

Oakhill, J., Cain, K., & Elbro, C. (2015). *Understanding and teaching reading comprehension: A handbook*. New York: Routledge.

Rapp, D.N., van den Broek, P., McMaster, K.L., Kendeou, P, & Espin, C.A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*(4), 289-312.

Raudenbush, S.W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd Ed.)*. Thousand Oaks: Sage.

Ritchey, K.D., & Speece, D.L. (2006). From letter names to word reading: The nascent role of sublexical fluency. *Contemporary Educational Psychology, 31,* 301-327.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies of causal effects. *Biometrika*, *70*, 41-55.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3-4), 169-188.

Smith S. B., Simmons, D. C., & Kame'enui, E. J. (1998). Phonological awareness: Instructional and curricular basics and implications. In D. C. Simmons & E. J. Kame'enui (eds.), *What reading research tells us about children with diverse learning needs: Bases and basics.* Mahwah, NJ: Lawrence Erlbaum Associates.

Soland, J. & Thum, Y.M. (2019). Effect Sizes for Measuring Student and School Growth in Achievement: In Search of Practical Significance (EdWorkingPaper No.19-60). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-60

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

Stuart, E.A., & Rubin, D.B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155-176). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/978141299562

Torgesen, J.K.,. Wagner, R.K., & Rashotte, C.A. (1999). Longitudinal studies of phonological processing and reading. *Journal of Learning Disabilities, 27*(5), 276-286.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. *What Works Clearinghouse Standards Handbook, Version 4.0.* Retreived from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition, 51*(1), 91-103.

Yuill, N. (2009). The relation between ambiguity understanding and metalinguistic discussion of joking riddles in good and poor comprehenders: Potential for intervention and possible processes of change. *First Language, 29*(1), 65-79.

Zipke, M., Ehri, L.C., & Smith Cairns, H. (2009). Using semantic ambiguity training to improve third graders' metalinguistic awareness and reading comprehension: An experimental study. *Reading Research Quarterly, 44*(3), 300-321.

Appendix

Table 1: Analytic Sample Descriptives

	Kindergarten		**Grade 1**		**Grade 2**	
	Treatment	Control	Treatment	Control	Treatment	Control
Charter	2.0%	2.0%	1.0%	3.0%	2.0%	2.0%
Female	48.0%	48.0%	48.0%	48.0%	49.0%	48.0%
Male	52.0%	52.0%	52.0%	52.0%	51.0%	52.0%
Is Esl	41.0%	39.0%	34.0%	32.0%	29.0%	30.0%
Is Special Ed	10.0%	11.0%	9.0%	11.0%	8.0%	12.0%
Magnet	10.0%	13.0%	12.0%	16.0%	14.0%	16.0%
African American	6.0%	9.0%	6.0%	9.0%	7.0%	9.0%
Hispanic	85.0%	79.0%	85.0%	79.0%	84.0%	80.0%
Other	5.0%	5.0%	4.0%	5.0%	4.0%	5.0%
White	4.0%	7.0%	4.0%	6.0%	5.0%	6.0%

Table 2: Composite Score Comparisons

		Raw N	**Weighted N**	**Mean MOY Composite Score**	**Mean EOY Composite Score**
Kindergarten	Comparison Group	18990	2105.0	130.9	140.2
	Tx Group	4647	4647.0	137.5	152.9
Grade 1	Comparison Group	15797	2552.8	160.7	163.4
	Tx Group	6041	6041.0	170.1	177.6
Grade 2	Comparison Group	17525	3466.2	205.7	236.7
	Tx Group	5981	5981.0	227.9	269.8

Table 3: All Score Comparisons

			Raw N	**Weighted N**	**Mean MOY Score**	**Mean EOY Score**
Kindergarten	Composite	Comparison	18990	2105.0	130.9	140.2

	Score	Group				
		Tx Group	4647	4647.0	137.5	152.9
	LNF	Comparison Group	18990	2105.0	38.6	51.3
		Tx Group	4647	4647.0	39.7	54.2
	NWF (CLS)	Comparison Group	18989	2105.0	27.9	44.2
		Tx Group	4647	4647.0	28.6	48.0
	PSF	Comparison Group	18990	2105.0	30.6	44.6
		Tx Group	4647	4647.0	32.9	50.8
Grade 1	Composite Score	Comparison Group	15797	2552.8	160.7	163.4
		Tx Group	6041	6041.0	170.1	177.6
	DORF (Accuracy)	Comparison Group	15797	2552.8	69.4	79.0
		Tx Group	6041	6041.0	74.0	85.2
	DORF (Fluency)	Comparison Group	15797	2552.8	32.8	51.3
		Tx Group	6041	6041.0	35.4	56.1
	NWF (CLS)	Comparison Group	15797	2552.8	58.7	73.4
		Tx Group	6041	6041.0	62.8	80.7
	NWF (WWR)	Comparison Group	15797	2552.8	17.6	23.1
		Tx Group	6041	6041.0	18.6	25.7
Grade 2	Composite Score	Comparison Group	17525	3466.2	205.7	236.7
		Tx Group	5981	5981.0	227.9	269.8
	DORF (Accuracy)	Comparison Group	17523	3464.8	86.7	88.7
		Tx Group	5981	5981.0	92.2	94.3
	DORF (Fluency)	Comparison Group	17523	3464.8	72.6	84.1
		Tx Group	5981	5981.0	79.6	94.6
	DORF (Retell Quality)	Comparison Group	14831	2819.3	2.6	3.0

		Tx Group	5376	5376.0	2.6	3.0

Table 4: Minigame Descriptions

Game	Grade	Skill
Zoom Boom	K	Phonological Awareness: Rhyming
Gem & Nye	K-1	Phonological Awareness: Blending
Word Bots	K	Phonological Awareness: Segmenting
All Aboard	K-1	Phonological Awareness: Segmenting
Cut it Out	K-1	Phonological Awareness: Phoneme Isolation
Picky Goblins	K-1	Phonics: Letter-Sound Correspondence
Hangry Goblins	K-2	Phonics: Letter-Sound Correspondence
Grumpy Goblins	1-2	Phonics: Letter Combinations
Rhyme Time	K-1	Phonics: Early Decoding
Word City	K-2	Phonics: Early Decoding
Tongue Twist	1	Phonics: Early Decoding
Food Truck	1-2	Phonics: Early Decoding
Sort it Out	1-2	Phonics: Early & Advanced Decoding
Curioso Crossing	K-2	Phonics: Early & Advanced Decoding
Read All About It	K-2	Phonics: Early & Advanced Decoding
Word Slide	1-2	Phonics: Advanced Decoding
Storyboard	K-2	Comprehension: Inferences
Super Match	1	Comprehension: Cognitive Flexibility
Show-Off	2	Comprehension: Text Structure
Because This, That	2	Comprehension: Text Structure
Connect It	2	Comprehension: Syntactic Awareness /

		Connectives
Umask That	2	Comprehension: Syntactic Awareness / Anaphora
Message in a Bottle	2	Comprehension: Syntactic Awareness
What’s the Big Idea	K-2	Comprehension: Main Idea
Story Box	K	Comprehension: Story Elements
Picture This	1-2	Comprehension: Story Elements
Best Buddy	1-2	Comprehension: Character Traits
Book Club	2	Comprehension: Compare & Contrast Texts
Tube Tales	2	Comprehension: Text Schema
Field Observer	2	Comprehension: Evidence for Inferences
Debate-a-ball	2	Comprehension: Claims & Evidence
Sloppy Scrolls	2	Comprehension: Comprehension Monitoring
Sticker Book	K	Vocabulary: Categories & Word Relationships
Word Raiders	1-2	Vocabulary: High-Utility Curricular Words
Shades of Meaning	1-2	Vocabulary: Word Relationships
Ink Blott	1-2	Vocabulary: Affixes
Punchline	2	Vocabulary: Multiple Meaning Words

Note: This is a comprehensive list of mini-games available for students in grades K-2. Not all games are played at all times of year, as this is dependent on a student’s placement and progress within Amplify Reading. For detailed descriptions of these games, visit https://amplify-com-mktg.imgix.net/app/uploads/2019/07/30161852/AmplifyReading_Program-Guide_07.19.19_Final-Draft.pdf