Realigning Early Reading Instruction with Research: A Preliminary Evaluation of Two Research-Based Early Reading Programs

by

Julia B. Lindsey

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Educational Studies)
in the University of Michigan
2021

Doctoral Committee:

Professor Nell K. Duke, Co-Chair
Associate Professor Christina J. Weiland, Co-Chair
Professor Brian A. Jacob
Professor Heidi Anne E. Mesmer, Virginia Polytechnic Institute and State University

Julia B. Lindsey

jblind@umich.edu

ORCID iD: 0000-0002-2468-7041

Dedication

Dedicated to all my students. I wish I knew then what I know now.

Acknowledgements

I am deeply grateful to everyone who supported me throughout this program and dissertation. First, I would like to thank my dissertation committee. Nell, I will never have enough words to thank you. In the fall of 2016, when I began applying to graduate programs, I was committed to two things: getting into a developmental psychology program and never returning to the Midwest. One phone call with you showed me my priorities were not quite right, and I've never been more thankful (though I could live without the Michigan winters). Thank you for bringing me into reading research, for believing in me, for teaching me, for standing up for me. Thank you for sharing your wisdom. I look forward to being your friend and colleague for many, many years to come.

Chris, thank you for welcoming me with open arms into your lab. You made me feel so included in a journey that can feel so lonesome. Thank you for all your support and expertise, even when my questions felt especially basic, you always helped. I'm also so grateful for how you have championed me and my work. Without pushes from you, I'm not sure this dissertation would look remotely like it does today. Heidi Anne, I am so grateful that you've shared your expertise with me, in this and in other work. It's been a delight to learn from you for many years. Brian, thank you so much for your voice, expertise, and commitment to rigor. To all of you, my many, heartfelt thanks for your time, expertise, and flexibility.

Second, I would like to thank my partners in this work. First, to the Boston Public Schools Department of Early Childhood and Brooke Childs. I'm so honored that you trusted me to help

your teachers and young readers. It has been a great honor to support your incredible work. Jason Sachs and Andrea Zayas, thank you for signing off on years of projects and being committed to improving reading. Brooke, I owe you a deep debt of gratitude and cannot thank you enough for everything. None of my work on decodable texts would exists without you. You have truly been my partner in this work and my friend.

To the Center for Black Educator Development and Jonah Edelman—another unspeakably large thank you. I'm humbled by your continued interest in working with me. Jonah, Sharif, Kelli, Shayna, and Makael, I thank you all for your support, for believing in me, and for being my friends. There are not words that describe my deep gratitude, humility, and pride for working with you and walking with you in the fight for equity and justice in the education of Black children.

Third, to Amanda Weissman, thank you for the hours and hours of support in statistics. I couldn't have done this without you.

Fourth, to my family. Mom and Dad, thank you for everything. I'm so happy I am your daughter. Caroline and David, thank you for setting a standard of excellence and always believing that I could reach it with you. And David, thank you for telling me I could do this Ph.D. and for constantly listening to me moan about statistics. Mitch and Dhvani, thank you for being a part of our family. I'm sorry you've had to listen to me yell about educational inequities for so many Christmases—to be honest, that will not stop once I graduate. To my grandparents, thank you for always telling me how proud you are of me. And one non-human thank-you to Kiwi, my dog, for keeping me sane during the COVID-19 pandemic. I love you all.

Finally, but certainly not least, I would like to thank my dear friends. Mary Taylor, Ariane, Olivia, Blake, and Kate, thank you for walking through life with me. I love you all so much more

than words can say. Meghan, Rachael, Hannah, Sarah, Maggie, Melanie, and Alessandra, I'm so glad I wasn't doing this alone. Your friendships mean the world to me. To the Duke advisees, the EEL Lab members, the Great First Eight team, and countless others who have supported and inspired me, thank you.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Early reading instruction is foundational to children's success in school and in life. Early reading abilities predict long-term reading and other academic outcomes (e.g., Cunningham & Stanovich, 1997; Hernandez, 2011; Juel, 1988). Recent estimates show that only thirty-five percent of U.S. fourth graders are proficient readers, and the proportion of children of color and children experiencing poverty who are proficient readers is even smaller (NAEP, 2019). Despite decades of research, policy action, and other attempts to improve early reading outcomes, the most common curricula and texts used in early reading instruction do not tend to be research-tested, nor are they necessarily reflective of literacy research (*EdWeek*, 2019; Hiebert, 2017; Simba Information, 2017). Developing more research-aligned curricula and texts to support children in early reading is utterly essential to increase the proportion of students experiencing success.

This dissertation consists of two stand-alone manuscripts that attempt to add to understandings, both in research and practice, about improving early reading instruction. Both papers are also related as preliminary attempts to estimate the impact of reading programs. The first paper presents a clustered observational study of the implementation of a new type of multiple criteria texts and accompanying instruction in first-grade classrooms in a large, metropolitan school district in the northeastern United States. The texts attended to children's learning of phonics and content, as well as attempting to privilege culturally relevant topics and characters. Unfortunately, due to the impact of COVID-19, only a fraction of the intervention was able to be implemented. Not surprisingly, then, compared to other schools in the district, the ten volunteer treatment schools' students did not have statistically significantly different word reading gains in the first half of first grade. Exploratory fidelity evidence suggests texts

implemented along with in-the-moment phonics-focused word reading instruction may have the potential to improve word reading outcomes. There is a need for continued future research of multiple criteria texts.

The second manuscript presents a pre/post-test study of a summer literacy program, the Freedom Schools Literacy Academy (FSLA), a summer program designed to support Black elementary schoolers' reading and racial identity development. This summer program combines research-based reading instruction within a culturally responsive framework. The preliminary evaluation of the virtual/distance-learning version of this program investigated the effects of the program for 83 children in listening comprehension, word reading, oral reading fluency, and racial attitudes. Results indicated that program participation resulted in statistically significant growth in all areas. Findings indicate the promise of this program, and, more broadly, the potential to support simultaneously support children's foundational reading abilities, comprehension, and development of a positive racial identity.

This dissertation preliminarily evaluates two new programs/interventions to support early reading. These papers add to the limited knowledge base on the intersection of culturally responsive practices and research-based early literacy instruction. Both studies, furthermore, supported the translation of research into practice by evaluating promising programs that remain in place, continue to be improved upon, and are ripe for further research in a post-pandemic context.

References

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*(6), 934-944.

EdWeek. (2020). *Early reading instruction: Results of a national survey.* Retrieved from https://epe.brightspotcdn.com/32/4f/f63866df760fb20af52754fd07ff/ed-week-reading-instruction-survey-report-final-1-24-20.pdf

Hernandez, D. J. (2011). *Double jeopardy: How third-grade reading skills and poverty influence high school graduation.* Annie E. Casey Foundation.

Hiebert, E. H. (2017). The texts of literacy instruction: Obstacles to or opportunities for educational equity? *Literacy Research: Theory, Method, and Practice*, *66*(1), 117-134.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of educational Psychology*, *80*(4), 437-447.

Simba Information (2017). *K-12 Reading market survey report 2017*. Simba Information.

Chapter 1 Introduction

An estimated thirty-four percent of U.S. fourth graders are reading at a "below basic" level as defined by the National Assessment of Education Progress Assessment (2019). Children who tested into this level likely had challenges locating relevant information to answer questions, making inferences, and identifying details to support conclusions on this assessment (based on NAEP's category definitions; NAEP, 2019). In addition to potential barriers to demonstrating comprehension of a text, children identified as reading at a "below basic" level also tend to read words in isolated and connected texts with statistically significantly less automaticity, accuracy, and expression than their more proficient peers (White et al., 2021). The magnitude of this problem cannot be overstated. The needs of one-third of U.S. children are not met in elementary school reading instruction.

For the past several decades there has been widespread recognition that early reading instruction needs to include systematic, explicit phonics instruction. Research has continually confirmed that phonics instruction supports the development of proficient word reading (e.g., de Graaff et al., 2009; Ehri et al., 2001; Henbest & Apel, 2017; Torgerson et al., 2018). This research has been codified by various entities, such as the Common Core State Standards (National Governors Association, 2010) and the Report of the National Reading Panel (2000). A recent survey of K-2 teachers revealed that 70% believe they put "a lot" of emphasis on phonics in their reading instruction (*EdWeek*, 2020). Despite the prominence of phonics, however, there

has been no statistically significant improvement in proficiency in fourth-grade reading on the National Assessment of Educational Progress[1] in the past 14 years (NAEP, 2019).

Why hasn't this emphasis on phonics, across stakeholders, translated into clear gains in reading achievement? While this question likely has a nuanced, debatable answer, one major contributor is that phonics instruction simply isn't addressing all of children's needs as emergent word readers.[2] Identifying instructional practices and programs that do address more of children's needs is critical for educators, researchers, policymakers, and funders. Phonics instruction in many classrooms tends to be only tangentially informed by research and is disconnected from other learning, reading, and cultural experiences. In this dissertation, I present two studies that attempted to address these challenges in early grade phonics instruction.

Overview of the Dissertation

This dissertation draws on theories and empirical work about word reading development, text supports for early readers, and culturally responsive, relevant, and sustaining practices, and uses quantitative methods. In this dissertation, I report the findings of two research studies using an alternative format that includes two journal-length manuscripts, prepared for submission to research journals. These manuscripts are written with all the typical components of a journal article. The alternative format may help the findings from these studies reach a broader audience than typically formatted dissertations (Duke & Beck, 1999).

---

[1] The NAEP assessment, a congressionally mandated assessment, has been administered to a national representative sample of fourth graders periodically since 1992 and every two years since 2003. In 2019, 150,600 fourth graders took the NAEP Reading assessment and 1,800 took the Oral Reading Fluency portion.

[2] Here, and in much of this work, I draw a distinction between word readers (someone who can automatically and accurately read words) and proficient reading more broadly in recognition that phonics alone is certainly not enough (and never has been enough) to lead to proficient reading comprehension.

In the first paper (*Redefining Research-Aligned Decodable Texts*), I estimate the impact of a multiple criteria text supplement to a first-grade phonics curriculum. This is the first study to investigate the use and impact of multiple criteria texts at a large scale. This study is a clustered observational study. I explore a recently developed technique for multilevel matching to estimate the impact of the treatment. In this study, I add to the literature about whether multiple-criteria texts support children's reading development by answering the following questions: First, do multiple-criteria, content-connected texts impact first-graders' word reading, as measured by NWEA Map Reading Fluency? Second, is higher fidelity to implementation positively associated with gains in first-graders' word reading?

The first study involved a sample of schools in the Boston Public Schools district. Ten schools volunteered to have first-grade classrooms participate by using multiple criteria text in instruction. I compared students' word reading gains in the first part of first grade (September-December) in treatment schools to matched pairs in other schools in the same district. Due to the COVID-19 pandemic, the study only includes information from the first portion of the school year.

Teachers at treatment schools received 40 "Beyond Decodable" texts (content-connected, multiple criteria texts) designed to support children's application of the district's phonics curriculum in connected text, while also supporting other aspects of children's reading. Most notably, the texts were linked to the district's knowledge-building curriculum ("Focus on First") and aimed to be culturally responsive to the learners in Boston. Treatment condition teachers also participated in a one-hour professional development session and had access to a website with additional information and lesson plans. This "light touch" intervention asked teachers to use up to 20 texts for at least one reading experience per week (small group, whole group, or

scaffolded independent reading) with a majority of their students over the first portion of this school year.

Results of a multilevel linear regression model with multilevel matches (to improve balance between conditions and, therefore, mimic a randomized control trial more closely) indicated no statistically significant differences in the word reading outcomes of first graders. Robustness checks indicate this result was not particularly sensitive to data analytic choices. Given the variation in observed fidelity to implementation, I also investigated the association of implementation (measured three ways) and word reading outcomes. Multilevel linear regression models again indicated no statistically significant differences in the word reading outcomes of first graders with teachers who appeared to implement the texts with greater fidelity compared to those with lesser fidelity, measured as a binary or as a scale. A multilevel linear regression model did indicate that students engaged in instruction about using phonics to read words while reading Beyond Decodables demonstrated marginally statistically significant gains compared to students in other treatment classrooms. Due to constraints of the sample size, this evidence is extremely preliminary, but does suggest that phonics-focused instruction plus multiple criteria text reading may support word reading outcomes. This study, encumbered by the COVID-19 pandemic, does demonstrate that multiple criteria texts can be meaningful and linked to content and that multiple criteria text reading in conjunction with phonics instruction has the potential to support word reading development. Findings from this study suggest that additional research is necessary to understand the impact of supplementing phonics instruction with multiple criteria texts.

The second study (*A Preliminary Evaluation of Freedom Schools Literacy Academy*) also investigated a program with research-based early reading instruction, multiple criteria texts (different texts and criteria than Study 1), and an emphasis on culturally responsive pedagogies.

In this study, I estimate the impact of a culturally responsive and sustaining summer literacy program on early elementary schoolers' literacy outcomes and racial attitudes. Due to the COVID-19 pandemic, the program followed a virtual/distance-learning format; thus, this study is unique in adding to understandings about literacy instruction in a virtual format.

Freedom Schools Literacy Academy was a 4-week program with approximately 30 hours of student-program contact. Children participated in explicit, systematic phonics lessons with connected multiple criteria texts, culturally responsive read-alouds, and a motivational experience each day. The program's teachers ("Servant Leader Apprentices") were college-age pre-service teachers who engaged in over 30 hours of professional development and coaching throughout the program.

I estimate the impact of the Freedom Schools Literacy Academy using a pre/post-test design with measures of listening comprehension, word reading, oral reading fluency, and racial attitudes. In this study, I add to the literature about how summer literacy programs can support children's learning and how early literacy instruction can support racial identities by answering the following questions: First, did children who participated in Freedom Schools Literacy Academy show gains on measures of listening comprehension, word reading, and positive racial identity? Second, are gains in word reading, listening comprehension, and positive racial identity associated with children's characteristics (age, gender, and socioeconomic status)?

The Freedom Schools Literacy Academy study involved a sample of 83 children who voluntarily attended the program in the summer of 2020. I compared children's pre-test scores (on the measures above) to their scores at the end of the program using t-tests. I then investigated the associations of children's characteristics and their gains in these areas with linear regression

models. In the second paper, I also discuss how the gains made by children in this program compare to gains reported in other research on summer literacy programs.

Results of the t-tests indicated statistically significant gains in children's listening comprehension, word recognition, oral reading fluency, and racial attitudes. Given the study's design and other constraints, these results should be interpreted with caution; however, results do point to the possibility of positively addressing foundational reading instruction and comprehension while supporting children's racial identity. To my knowledge, this is the first study to examine early elementary literacy instruction with these three components using quantitative methods. Findings from this study further suggest that children are able to make gains in literacy in a virtual learning environment.

In summary, these two dissertation studies, although limited by methodological challenges and the impacts of the COVID-19 pandemic, point to the positive possibilities of attending to children's needs in foundational reading while also supporting their identities and knowledges. When research-based teaching techniques are used, programs and curricula can integrate deliberate attention to children's racial, ethnic, and/or cultural identities without sacrificing, and possibly bolstering, children's foundational reading abilities.

References

de Graaff, S., Bosman, A. M., Hasselman, F., & Verhoeven, L. (2009). Benefits of systematic phonics instruction. *Scientific Studies of Reading*, *13*(4), 318-333.

Duke, N. K., & Beck, S. W. (1999). Education should consider alternative formats for the dissertation. *Educational Researcher*, *28*(3), 31–36.

EdWeek (2020). *Early reading instruction: Results of a national survey.* Retrieved from https://epe.brightspotcdn.com/32/4f/f63866df760fb20af52754fd07ff/ed-week-reading-instruction-survey-report-final-1-24-20.pdf

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, *71*(3), 393-447.

Henbest, V. S., & Apel, K. (2017). Effective word reading instruction: What does the evidence tell us? *Communication Disorders Quarterly*, *39*(1), 303-311.

NAEP Reading Report Card. (2019). Retrieved from https://nces.ed.gov/nationsreportcard/

Torgerson, C., Brooks, G., Gascoine, L., & Higgins, S. (2018). Phonics: reading policy and the evidence of effectiveness from a systematic 'tertiary' review. *Research Papers in Education*, *34*(2), 208-238.

White, S., Sabatini, J., Park, B. J., Chen, J., Bernstein, J., & Li, M. (2021). *The 2018 NAEP oral reading fluency study.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, Assessment Division. Retrieved from https://nces.ed.gov/nationsreportcard/subject/studies/pdf/2021025_2018_orf_study.pdf

Chapter 2 Redefining Research-Aligned Multiple Criteria Texts: A Preliminary Evaluation of a Light-Touch Implementation of Content-Connected, Multiple-Criteria Texts in First Grade

Abstract

This study investigated the estimated impact of a multiple criteria text supplement to a first-grade phonics curriculum on children's word reading outcomes. In this clustered observational study, twenty-five first-grade teachers at ten schools implemented a series of multiple criteria, content-connected texts during the first half of the school year. These texts were written to support children's ability to apply phonics knowledge in meaningful texts, related to content learning and children's background knowledge. Using multilevel matching to construct an appropriate counterfactual group, results did not indicate a significant impact of reading these texts. Exploratory fidelity evidence suggests that texts implemented along with in-the-moment phonics-focused word reading instruction may support word reading outcomes. Findings, however, are complicated by limitations of methodology, implementation fidelity, and the COVID-19 pandemic.

Educators, policymakers, and researchers have long discussed the best types of texts to support beginning readers (e.g., Gourley, 1984; Hiebert & Mesmer, 2013; Mesmer et al., 2012; Tortorelli, 2019). Prolific independent reading in the early years is a uniquely powerful predictor of reading achievement (e.g., Cunningham & Stanovich, 1997; Sparks & Murdoch, 2014). Thus, it is essential to ensure that young readers can experience success early in their literacy development. Texts for early readers matched to readers' developing skills and needs to scaffold and support reading development may support their ability to experience success. When children read in appropriately challenging, engaging texts, they have the chance to develop the skillset and knowledge of proficient readers.

Despite the importance of appropriate texts, research has yet to discern what combination of elements of texts are most supportive for readers at particular stages of development. Decodable texts (defined in this study as texts with a high proportion of words with graphophonemic patterns and high frequency words known to the reader) are one text type that some research suggests may appropriately support beginning readers above and beyond other text types (e.g., Allor et al., 2020; Beverly et al., 2009; Cheatham et al., 2014; Chu & Chen, 2014; Compton et al., 2004; Hiebert & Fischer, 2007, 2016; Juel & Roper-Schneider, 1985; Menon & Hiebert, 2005; Mesmer, 2005, 2010). Decodable texts, and more recently, multiple criteria texts, complement systematic, explicit instruction in phonics by giving children the chance to practice using grapheme-phoneme correspondences in reading actual texts to support orthographic mapping and automatic word recognition (see The Development of Accurate Word Reading below). Today there is little dissent in the research community that word recognition is letter-based, strengthening the decades-long theoretical argument for decodable texts (e.g., Grainger, 2018). Opportunities to decode words supports automatic word recognition, which

supports fluent reading and reading comprehension (Caravolas et al., 2019; Lepola et al., 2016; Hulme et al., 2015; Riedel, 2007; Roehrig el al., 2008; Torppa et al., 2016). Further, giving children the chance to contextualize and generalize phonics skills in texts is likely critical to reading success (Rupley et al., 2009; Stein et al., 1999; Taylor et al., 2000).

At present, decodable texts are fairly common (Mesmer, 2006), advocated for (such as in the reporting of Emily Hanford), and expected in new phonics curricula (see the requirements of the curriculum reviewing group, EdReports). The research, however, is not unanimously in favor of decodable text reading above and beyond other types of texts. Many of the studies on decodable texts are primarily observations of students' reading (e.g., Compton et al., 2004; Hiebert & Fisher, 2007; Mesmer, 2010). Others find some evidence in favor of decodable texts, but not necessarily for the entire sample or for each outcome measured (e.g., Beverly et al., 2009; Cheatham et al., 2014; Chu & Chen, 2014). Still other experimental and quasi-experimental work has failed to find evidence that decodable text reading supports beginning readers more than reading other text types (e.g., Jenkins et al., 2004; Price-Mohr & Price, 2018).

There are several possible explanations for these inconsistent findings, including challenges in implementation, differences in definitions and associated pedagogies, teacher implementation, dosage, and methodological choices. Another possible primary reason for this mix of findings may be the quality of the decodable texts themselves. Typical decodable texts may not be supportive enough of some aspects of reading development as they are often perceived as meaningless or unrelated to children's knowledge (Castles et al., 2018). These texts are often deemed so "restricted in word choice and so may tend to be inferior to real books in (a) maintaining children's interest and motivation to read and (b) in achieving the broader goals of building children's vocabularies and knowledge" (Castles et al., 2018, p. 16). Thus, even when

children can practice decoding in decodable texts, they cannot necessarily practice or apply the plethora of additional skills and knowledge proficient readers use within texts, such as developing knowledge and monitoring comprehension, nor do they necessarily support children's motivation to become extensive readers.

Indeed, typical decodable texts may not even support children in decoding (using knowledge of grapheme-phoneme relationships to recognize a word), as one estimate found that only between four and 21 percent of the words in so-called decodable books were actually decodable, based on each publisher's phonics program (Stein et al., 1999). Further, a more recent review of a popular reading intervention program (*My Sidewalks*) featuring decodable texts found that texts, on average, contained words that matched phonics instruction only 68% of the time (Murray et al., 2014). Most so-called decodable texts available for purchase do not provide the purchaser with sufficient information about the level of decodability in the texts for a particular point in time, leaving teachers to figure it out for themselves, a task that many teachers may not be prepared to accomplish.

In practice, there are even more barriers to acquiring high quality decodable texts and using them appropriately. First, as indicated above, it is challenging to find even moderate quality texts that are decodable or multiple criteria. Second, these texts tend to be expensive. For example, one complete classroom set (1 copy of each of 16 texts per 18 students) of decodable texts for first-graders from a popular phonics program retails for $2,563.20. Third, supporting a transition to using decodable texts may represent a philosophical paradigm shift for many teachers. Without acknowledging and supporting teacher's understandings of these texts, teachers may use decodable texts in ways that do not ultimately support children's word reading.

Given (1) the strong theoretical basis that children may be supported in word reading development by reading books with a high proportion of words they can decode, (2) the limited, but intriguing, empirical research offering preliminary evidence in favor of decodable texts, at least for some students, (3) the immense popular support for phonics-oriented reading reform, and (4) the practical challenges in implementing these texts in classrooms, it is necessary to continue to research texts with a high level of decodability. It is, however, also necessary to acknowledge the lack of consistent impact of these texts and continue the search for the most optimal text features for beginning readers (at present, texts that privilege decodability and other factors critical for reading success are called "multiple criteria texts").

In this article, I begin by overviewing research to find these potentially optimal text criteria (to create multiple criteria texts) by discussing the development and instruction of word reading, decodability, and other features related to accurate, automatic word recognition, and features to support children's comprehension and knowledge. Then, I describe a quasi-experimental evaluation of the implementation of a series of multiple-criteria, content-connected texts ("Beyond Decodables") and associated resources and training in first-grade classrooms, the impact of the offer of these texts on word reading outcomes, and variability in teacher implementation and children's associated word reading outcomes. Finally, I discuss the implications and limitations of this research.

Review of Literature

The Development of Accurate Word Reading

The ability to automatically recognize and read a large number of words is a critical component of fluent reading and the goal of most foundational skills instruction due to its relation to later overall reading fluency and comprehension. Empirical research has also found

that, for beginning readers, oral reading fluency within a text is related to comprehension of the same text and that decoding accuracy may be particularly related to comprehension (Juel et al., 1986). Longitudinal studies continually find that letter-sound knowledge predicts decoding skills, which predicts reading fluency and comprehension (Caravolas et al., 2019; Lepola et al., 2016; Hulme & Snowling., 2015; Riedel, 2007; Roehrig el al., 2008; Torppa et al., 2016). Although successful comprehension of a text requires more than just accurate word reading, research and theory indicate that accurate oral word reading supports successful comprehension (e.g., Amendum et al., 2018; Riedel, 2007; Roehrig et al., 2008).

Substantial research demonstrates that word reading can be efficiently and effectively taught in part through explicit instruction in grapheme-phoneme relationships in an evidence-based and logical order (i.e., explicit, systematic phonics) (de Graaff et al., 2009; Henbest & Apel, 2017; Torgerson et al., 2018). Ideally, this instruction supports an eventual ability to read mostly by automatic recognition of words' pronunciations and meanings from memory (Ehri, 2005). Proficient, fluent readers are able to read mostly by automatic, accurate recognition of words' pronunciations and meanings from memory (Ehri, 2005).

So how do children likely learn to automatically recognize words and become fluent readers who can comprehend complex texts? To read an individual word, a child needs to link the word's orthographic information (spelling) to its phonology (pronunciation) and sematic information (meaning). This is called an *orthographic map* (see Figure 1). When a reader has an orthographic map of a word, the word's meaning and pronunciation are automatically, effortlessly retrieved from memory when the reader encounters the word (Ehri, 2005, 2014, 2020). This process is often described via Share's (1995) self-teaching model, in which successful encounter(s) with a word allow a reader to acquire new word-specific orthographic

information. To build orthographic maps and their sight word vocabulary, readers need to form connections between spellings and pronunciations. This is achieved through decoding (Ehri, 2014; Share, 2004). When a child has the opportunity to decode a word several times (estimates range from one to eight times on average, likely depending on the word, context, and child's knowledge; Bowey & Muller, 2005; Nation et al., 2007; Share, 2004), they can store this orthographic map in memory (often called *sight words*). Critically, knowledge and repeated exposure to letter strings and patterns in words seems to facilitate the generation of orthographic maps for novel words with the same orthographic structures, making this process much faster than simply memorizing every word in the English language (Cunningham et al., 2002; Hiebert & Fisher, 2016; Share, 2004).

In sum, the current research on the development of accurate word reading suggests that three instructional moves are critical. First, phonics instruction must be provided. Phonics instruction supports a beginning reader by providing a set of skills and knowledge in order to recognize words and create orthographic maps. Without phonics instruction, a child would need to rely on less efficient means to make connections between spellings and pronunciations and some children struggle to ever do so. Second, children need instruction in how to sound out words. In other words, young readers are best supported by explicit instruction in how to use grapheme-phoneme relationships to figure out a words' pronunciation. Third, in order to facilitate word reading, children need the opportunity to actually decode words. Children need many opportunities, likely in decontextualized and contextualized practice, to actively apply their knowledge of the linkages between spellings and pronunciations to build their sight word vocabulary.

Decodability Criteria

Theoretically, decodable texts support word reading development by providing an opportunity for children to apply phonics skills in context, which, over time, would allow children to improve their automatic word recognition. The potential of these texts, therefore, cannot be separated from the connected phonics curriculum and the reading context or instruction surrounding each text (e.g., Juel & Roper-Schneider, 1985). Thus, most studies of decodable texts are generally investigating the impact of decodable texts and a connected phonics program (e.g., Allor et al., 2020; Beverly et al., 2009; Cheatham et al., 2014; Chu & Chen, 2014; Compton et al., 2004; Hiebert & Fischer, 2007, 2016; Juel & Roper-Schneider, 1985; Menon & Hiebert, 2005; Mesmer, 2005, 2010). Research that informs our understanding of decodable texts typically either describes children's reading behaviors in texts with high proportions of decodable words or compares children's literacy outcomes after reading more or less decodable texts.

Observations of Children's Reading in Decodable Texts

Descriptive studies of children's reading behaviors in texts with a high percentage of decodable and high frequency words suggest that these texts may support more accurate, automatic reading, likely because children are able to apply knowledge and skills from phonics instruction in texts. Compton and colleagues (2004) investigated the reading behaviors of 248 second-grade children in a grade-level passage once per week across 15 weeks of instruction. They found that the percentage of high frequency words was significantly associated with reading accuracy and rate for all readers and the percentage of decodable words (defined as regular single syllable words) was associated with reading rate for readers identified as average achieving decoders (the percentage of decodable words accounted for 23% of variance in reading rate). Hiebert and Fisher (2007), in investigating a new measure of text difficulty, the Critical

Word Factor, also found that first graders' accuracy and rate were significantly better in books with fewer "hard words" (hard words defined as words that were not high frequency or decodable words). Mesmer (2010), in a study examining 74 first graders' accuracy and rate across the year in decodable versus qualitatively leveled texts, found that children's reading rate was higher in texts with a higher proportion of high frequency words (in this case, the qualitatively leveled texts). In the same study, for children reading a set of books tightly matched to their demonstrated phonics abilities, children's accuracy was significantly higher when reading decodable texts. However, studies did not examine whether students' more fluent reading of these texts resulted in more accelerated reading development.

Broadly, observations of first and second graders' reading in texts with a high proportion of decodable and high frequency words find that children tend to be more accurate and automatic readers than when reading other texts (Compton et al., 2004; Hiebert & Fisher, 2007), especially when the texts are directly matched to children's phonics instruction (Mesmer, 2005, 2010). These mostly small, observational studies offer insights into how children may benefit from decodable texts and support the additional findings of how reading decodable texts may support reading in other texts and development.

Evidence of Impacts on Children's Reading of Decodable Texts

The earliest study investigating the impact of decodability on early reading found that reading texts with more decodable words supported word reading growth for first-grade students better than reading less decodable texts (Juel & Roper-Schneider, 1985). In Juel and Roper-Schneider's (1985) experimental study, children were randomly assigned to read more or less decodable texts. All participating children's teachers transitioned from basal-focused reading instruction (defined in this report of research as reading instruction linked to a basal reader

series, such as *Dick and Jane*, and incidental phonics instruction) to an explicit, systematic phonics program plus reading groups. Ninety-three first graders in 11 classrooms at three schools were assigned to one of two reading groups, which were equivalent in baseline characteristics; one, with decodable texts, and two, with basal readers focused on high frequency words and with a higher proportion of non-decodable (at this stage) words. Although children developed similar letter-sound correspondence knowledge, children in the condition that emphasized decodable and regular words at the beginning of the year demonstrated superior skill at blending and reading novel words. Furthermore, evidence throughout the year indicated that children reading texts with more regular, decodable words were using letter-sound knowledge to identify words more than their peers, who primarily used guessing or context-based strategies. The authors postulate this may have facilitated word recognition development.

This study is important for three reasons. First, it set the stage for more studies investigating the impact of reading decodable texts by indicating that these texts support word reading development. Second, it demonstrated that children's early experiences and instruction in texts may lead to the use of more or less effective word recognition strategies. Third, as the authors controlled the entire reading instruction block, this study may be one of the most accurate depictions of the contrast in outcomes due to differences in text type, as all other studies reviewed above implemented supplemental elements rather than changing the entire reading block.

Another study that indicates decodable texts may support reading development describes children's application of phonics skills in decodable texts. Mesmer (2005) studied the impact of reading decodable texts matched to a coordinated phonics program on a small sample of first-grade readers. Twenty-three first graders, of average ability, participated in supplemental small-

group phonics lessons and were randomly assigned to read more or less decodable texts. Students reading decodable texts were more accurate in their word reading ($d = 0.87$) and applied their letter-sound knowledge more frequently ($d = 1.00$) than the comparison group. They also relied less on asking the examiner for pronunciations. Mesmer (2005) concluded that decodable texts gave students the opportunity to practice new phonics skills at a higher rate than traditional books, especially when the text is highly matched to the phonics lesson's content (Lesson-to-text-match, LTTM). For example, students might learn the digraph *sh* in a phonics lesson and then practice applying their new knowledge in a decodable book that repeats the *sh* pattern (e.g., *ship, she, shark*).

Two studies suggest the greatest benefit of decodable text reading for children early in reading development. Beverly, Giles, and Buck (2009) and Cheatham, Allor, and Roberts (2014) found that the impact of texts with a high proportion of decodable words on word reading outcomes may be moderated by students' initial reading abilities. Beverly and colleagues (2009) compared 32 first graders receiving three instructional options in small groups: (a) phonics plus decodable text reading; (b) phonics plus an authentic literature read aloud; and (c) listening to an authentic literature read aloud. First graders were randomly assigned to groups with blocking for disability status. These small groups were supplemental to typical instruction, which the authors do not describe. Results of this study are limited by the small sample size and lack of statistical power. All groups made gains in all measures of word reading and reading fluency. Below average readers who participated in the phonics plus decodable text reading condition had statistically significant gains in comprehension compared to average readers in the same condition, which the authors hypothesize supports the use of decodable texts with below average readers. In this condition, children did not receive comprehension instruction, nor were they

expected to make gains in comprehension; thus, this result suggests a connection between decodable text reading and comprehension, perhaps through the path to fluency described early in this paper. Although this result requires further investigation in larger, more controlled studies, it indicates that decodable texts may support reading outcomes, including comprehension, for certain children.

A second study also found that beginning decoders may benefit from reading decodable texts. Cheatham and colleagues (2014) investigated the impact of multiple-criteria texts (in this study, these texts had a high proportion of decodable words, matched to lessons, had a high proportion of high frequency words, and were written with substantial attention to meaningfulness) on second-grade reading outcomes. Acknowledging the issues with decodable texts, such as their tendency to be have limited meaningfulness and lack of attention to other word-level factors that may support beginning readers, authors such as Cheatham and colleagues (2014) have investigated the use of multiple criteria texts, or texts that are decodable and designed to incorporate other features that contribute to reading development. Sixty-two second graders were randomly assigned to read multiple-criteria texts or authentic literature during independent reading. Though multiple-criteria texts did not measurably improve reading outcomes for all second-graders, there was a moderate effect in word reading fluency for beginning decoders ($d = 0.67$). The authors concluded that the use of multiple-criterion texts supported beginning decoders in word reading, but may not be more supportive than other texts for typical second-grade readers (Cheatham et al., 2014). This study and the Beverly and colleagues (2009) indicate that children who are early in the development of decoding skills may be most likely to be supported by decodable texts, pointing to the potential impact in kindergarten and first grade, as well as for children working on early decoding skills in other

grades.

Another population that may particularly benefit from reading multiple criteria texts is English Language Learners. Chu and Chen (2016) randomly assigned four classrooms of Taiwanese second graders (n=117) to phonics teaching with or without meaningful and decodable multiple criteria texts. Much like many studies on decodable texts, this study is limited in statistical power, so results must be interpreted cautiously. Although both groups improved in measures of English word recognition and word reading in context, there was a statistically significant effect of the phonics plus multiple criteria text group in delayed post-test on word reading (d = 0.37). In other words, children who read multiple criteria texts along with receiving explicit phonics instruction continued to improve in word reading for at least two weeks after the intervention while children who did not read multiple criteria texts, but did have explicit phonics instruction did not continue to improve and indeed declined in word reading in the post-intervention period. Hiebert and Fisher (2016) similarly found that English Language Learners (n = 81) randomly assigned to supplemental decodable text intervention outperformed children in the control condition in reading fluency (improving an average of approximately 7 more words correct per minute). As large proportions of children in US schools are English Language Learners (ELLs), these findings provide preliminary support for widespread use in districts with high proportions of multilingual learners.

In oft-presumed contrast to the above studies, Jenkins and colleagues (2004) found no impact of decodable text reading on first-grade reading. Researchers (2004) randomly assigned 121 “at-risk” first-grade students from eleven urban public schools to one of three supplemental tutoring conditions: (a) business-as-usual control; (b) phonics plus reading texts that were more decodable; or (c) phonics plus reading texts that were less decodable (quasi-random assignment

with some limitations based on classroom schedules). Although both supplemental groups statistically significantly improved reading outcomes over the control group, they found no statistically significant differences in reading outcomes between the two text types (on decoding, word reading, spelling, reading comprehension, fluency, or reading in context). Although this finding appears to be in contrast to Juel and Roper-Schneider (1985), two methodological differences indicate the take-away is less clear. First, there is a potential lack of treatment-control contrast between the more and less decodable groups. In both conditions, children's teachers were likely to report using decodable texts during non-supplemental instruction (69.2% in more decodable versus 55.3% in less decodable group). Texts in both conditions contained similar proportions of high frequency words (27.1% in more decodable versus 30.9% in less decodable group). Additionally, as the year progressed, the difference in the more and less decodable texts decreased from 74% decodability difference to 11% decodability difference. Thus, the texts may have been offering similar benefits, especially toward the end of the year. Second, the authors postulate that tutors may scaffold book reading experiences enough, mitigating the potential impact of more decodable texts (which also scaffold a student's reading experience). Additionally, this study does not suggest that reading more decodable texts is *less* effective than reading other types of texts as both phonics plus reading conditions had statistically significantly improved reading outcomes compared to the control group (*d* ranging from 0.38-1.13 on eight measures of decoding, word reading, and reading comprehension).

One additional recent study did not find that reading decodable texts positively impacted children's reading development. Price-Mohr and Price (2018), in a very small (n =12), non-randomized study claimed that the four kindergarten-aged children in the less decodable texts group appeared to have greater gains in word reading than the four students in the more

decodable text group, although children reading both text types still outperformed the four students in the business-as-usual group. In both Jenkins and colleagues (2004) and Price-Mohr and Price (2018), children did demonstrate improvement beyond control conditions when a supplemental phonics and reading program was added, regardless of the text type. The particularly small number of children per condition in the Price-Mohr and Price (2018) study, combined with the similarity of the conditions and impact of the more and less decodable texts conditions in the Jenkins and colleagues' (2004) study, considered along with the findings of other studies reviewed suggest the need for continued study of decodable texts

Overall, the weight of the evidence of the data shows that decodable and multiple criteria text reading, in conjunction with systematic, explicit phonics instruction, is likely to support beginning decoders in applying phonics to word reading. One study (Beverly et al., 2009) further suggests that decodable text reading may support beginning decoders in reading comprehension gains. Although the body of research on decodable texts is very small and tends to be limited in statistical power, studies tend to point in a direction that suggests decodable texts, when written and implemented well, could live up to their theoretical promise. The somewhat inconsistence of results and limited number of studies, however, also suggests that research needs to continue to proceed with careful attention to a wider range of factors that may support early readers, beyond just text decodability.

Word-Level Criteria Beyond Decodability

The above research suggests the promise of—as well as the need for further research on—controlling texts based on the proportion of decodable and high frequency words as it relates to children's fluency and continued reading development. In this section, I considered the research on additional criteria in order to more fully support children's word reading. Two

additional word-level criteria that research suggests may support automatic, accurate word reading in texts are repetition and what I term, "word understandability." Much like research on the impact of decodable texts on reading behaviors and outcomes, there is, as of this writing, only limited research with relatively small samples of children to draw on; however, the research that is available indicates that considering word-level criteria beyond just decodability may be beneficial to children's reading.

Word repetition may contribute to self-learning of novel words (Share, 1995), which would contribute to eventual fluent reading (as detailed in the section, The Development of Accurate Word Reading**)**. One theory of automatic word recognition, the instance theory of automaticity (Logan, 1997), suggests that many repetitions, of words, letter combinations, and letters, are required for automaticity. Empirical research also suggests the repetition of words in text may improve automaticity. In a small experiment testing 42 elementary children's self-learning of novel words, Nation and colleagues (2007) found that increased exposure to a nonword across a text or set of flashcards improved orthographic learning, regardless of the word's context ($d = 0.60$, comparing one versus four exposures). This study suggests that repetition of unknown words in a text may support orthographic learning of the word, and, therefore, may support automatic recognition of that word. Research also suggests that the repetition of patterns in words may improve fluency. In another experiment with 81 first graders randomly assigned to read texts controlled for various word-level factors, Hiebert and Fisher (2016) found that reading texts containing words with a high proportion of repeated rimes improved reading fluency more than reading other decodable text types or authentic literature (with children improving, on average, by 2.8 words correct per minute each week, improving far more than the typical growth for similar students; Hasbrouck & Tindal, 2017). Word and letter-

pattern repetition within texts may support fluency, in addition to the support provided by decodability.

"Word understandability" is a term I use here to represent words that are imageable, concrete, and/or familiar. All of these concepts are closely linked, and, both individually and together, may support word reading accuracy. I am introducing this term merely for ease of describing these aspects of word-level semantic complexity and not to insinuate that these terms should be considered one construct. Word imageability is the degree to which a word evokes a particular mental image in a reader. Some research suggests that highly imageable words may be read faster by young readers, which could improve accurate word reading with in a text (Hargis & Gickling, 1978). In the same vein, in a study of second grade informational text complexity analyzing over 5,000 children's oral reading, Tortorelli (2019) found that word concreteness (the extent to which a word represents a tangible concept) uniquely explained variance in oral reading rate among factors of text complexity. Further, word concreteness explained more variance than other measures of semantic complexity (Tortorelli, 2019). Word familiarity refers to a readers' familiarity with both the visual form and meaning of a word. When readers are more familiar with a word, they recognize and read the word faster (Williams & Morris, 2004). These findings indicate that including highly familiar, imageable, and concrete words in texts, therefore, may improve word reading accuracy, even when these words are not highly frequent or decodable. In the case of first-grade readers, words that fall into all of these categories (such as *animal* and *friend*) may support word reading accuracy and oral reading rate.

Text-Level Criteria Beyond Decodability

Although reading accurately is part of proficient reading, the ultimate goal of reading is comprehension. As noted earlier, a major critique of decodable texts is that they are often

meaningless, and, therefore, may not contribute to children's growth in comprehension or knowledge development (Castles et al., 2018). At present, research has not examined the degree of meaningfulness or comprehensibility of decodable texts. Given the importance of reading comprehensible texts (discussed below), however, this critique is worthy of attention.

In developing additional criteria beyond decodability, it is important to investigate factors related to children's comprehension and knowledge. Reading meaningful texts across multiple genres is essential for developing proficient comprehension skills. Research suggests that when children's knowledge is reflected in texts, they read and comprehend texts better. In a study on word identification, Priebe, Keenan, and Miller (2012), in a study with 60 fourth-grade children, found that prior knowledge of a text's topic improved automaticity and accuracy of word reading, indicating that prior knowledge of a topic improves word reading. This word reading improvement may indirectly also improve comprehension (through the mechanisms discussed above). Other research indicates that having some prior knowledge of a topic enables greater comprehension (e.g., Bell & Clark, 1998; McNamara & Kintsch, 1996). More specifically, research further indicates that text topics that are not only reflective of prior knowledge but are also reflective of cultural knowledge may support improved comprehension. In one experimental study with 109 children, Bell and Clark (1998) found that Black children had improved recall and listening comprehension about a text that focused on Black characters and African American cultural themes compared to texts with White characters and Euro-American themes. Overall, research across several decades in both children and adults suggests that higher knowledge of a topic improves text comprehension.

Additional research indicates that children's background knowledge of a particular topic or domain may support text comprehension and knowledge development. In a large-scale study

with data from the Early Childhood Longitudinal Study–Kindergarten cohort and 13,292 children, Hwang and Duke (2020) examined the role of science knowledge in third-grade reading and found that topical science knowledge was associated with higher reading comprehension for students who are monolingual and students who are English learners. Research also suggests that texts can support children's knowledge development. In an experiment with 59 fourth-grade children randomly assigned to read set of texts that were conceptually coherent or not, Cervetti and colleagues (2016) found that reading a set of conceptually coherent texts (texts about one topic) supported topical knowledge and vocabulary development over reading a set of six informational texts each on different topics. This research, although admittedly with older students than the present study, suggests that young children exposed to sets of conceptually related texts may build content knowledge simply through reading. This research suggests that the cohesiveness within a text and across a text set may support young readers in developing knowledge about a topic.

Research further suggests that when a student is interested in a text, they may experience superior word reading and comprehension. Texts matched to students' knowledge and interest may help students persist in challenge tasks (Fulmer & Frijters, 2011), such as the challenge of learning to read by applying graphophonemic knowledge. In Fulmer and Frijters' (2011) study with 56 elementary schoolers, children who read a story they rated as interesting were almost twice as likely to persist in reading, even in highly challenging texts. Research also suggests that interesting topics for reading increase student interest and motivation (Schiefele, 1999), and that even in early elementary school, intrinsic reading motivation contributes to comprehension (Schiefele et al., 2016).

Collectively, this research suggests that texts that support children in engaging in meaningful, conceptually coherent texts matched to children's individual background knowledge and interests may support additional aspects of reading development typically neglected in the development of decodable texts.

The Current Study

Research on what may support beginning readers centers on criteria that may address a range of student needs, including word reading and fluency, comprehension, and knowledge-building content. The evidence clearly indicates that simply controlling a text for its decodability and proportion of high frequency words may not lead to the creation of texts that support children fully in word reading nor are these texts likely to foster children's engagement and knowledge development.

The texts created for the present study are multiple criteria texts. They were designed to support first-grade readers in word reading by being highly matched to phonics instruction (based on decodability and high- frequency words) and including highly imageable and familiar vocabulary. The texts were designed to support readers in a motivating and content-connected knowledge building curriculum by being related to children's prior knowledge and based on conceptually coherent topics from literacy, science, and social studies instruction. The texts also aimed to be culturally relevant and place-based by featuring settings, characters, languages, and experiences that reflect the culture, practices, and experiences of the children in the study. Finally, acknowledging the importance of the instruction connected to these texts, the texts were not implemented alone; all teachers also received professional development and lesson planning materials matched to each text to support high-quality instruction.

In this study, I add to the literature about how multiple-criteria (and decodable) texts support children's reading development by answering the following questions:

1. Do multiple-criteria, content-connected texts impact first-graders' word reading, as measured by NWEA Map Reading Fluency?
2. Is higher fidelity to implementation positively associated with gains in first-graders' word reading?

The findings from this quasi-experimental evaluation of the implementation of a series of texts ("Beyond Decodables") in first-grade classrooms contribute to multiple literatures, policy, and practice. First, to my knowledge, this study is the first to employ a "light-touch" implementation of texts with a high level of decodability across a large sample of children, teachers, and schools, offering a policy-oriented understanding of the impact of texts with limited professional development, time, and money. Second, this is the first study to examine texts with these specific criteria and may help researchers and practitioners consider what criteria to include in other new texts. Third, this study serves as a pilot for future studies unencumbered by the impact of the COVID-19 pandemic. Fourth, this study using a relatively new procedure for clustered observational studies that allows for multilevel matching, adding to the literature on the utility of this technique.

Methods

To address the research questions, I developed a series of multiple-criteria, content-connected texts along with lesson planning materials and professional development. This study investigates the estimated impact of an offer of supplemental content-connected decodable texts, related materials, and professional development. In this clustered observational study, I tested the impact of this offer using multilevel matching to construct a comparison group and multilevel

linear regression. I also examined the association of the intervention's impact with implementation fidelity.

Beyond Decodables Texts

The Beyond Decodables supplement consisted of an offer of 40 multiple criteria texts created by the author of this paper (see Appendix 1.A for an example text and lesson plan), each matched to a specific week of content in the district's first-grade curriculum, *Focus on First*, and phonics program (Wilson Fundations®). Due to the impacts of the COVID-19 pandemic, teachers were only able to use at most half of these texts before the second test date (the third test date was cancelled due to the pandemic). Texts, materials, and professional development were based on learnings from a pilot study during the previous (2018-2019) school year. Additionally, all texts were reviewed by a team of six teachers and district staff for alignment to curriculum, place-based appropriateness, cultural relatedness, and engagement for children. Beyond Decodables were designed to address issues with current decodable texts and provide students a more supportive text environment.

In writing Beyond Decodables I attended to several major criteria (all reviewed above) to influence children's word recognition, ability to understand a word's meaning, and comprehension of the text as a whole (see Figure 2). First, there were four word-level criteria: (a) word decodability (in general and in terms of LTTM); (b) high frequency words; (c) word repetitions; and (d) word understandability of words deemed not-yet-decodable. Second, there were three text-level criteria: (a) conceptual coherent and related to science or social studies; (b) matched to children's perceived background knowledge, with culturally responsive and place-based topics, characters, and settings; and (c) enjoyable. See Appendix 1.B for descriptions of each criterion in each text.

Word-Level Criteria

Decodability. Each text was 80% decodable based on its particular placement in the scope and sequence of the phonics program, Fundations®, the district-adopted program. Fundations® is a commonly used phonics program across the country and is used in at least two other large metropolitan school districts. A word was considered decodable if it only contained grapheme-phoneme relationships previously introduced in the phonics program's scope and sequence or if a word was a high frequency word previously introduced in the phonics program's scope and sequence. Although there is not consensus about the minimum percentage of decodable words necessary to be a decodable text (Cheatham et al., 2014), several state laws (Foorman, Francis, Davidson, Harm, & Griffin, 2004) require texts to be 75-80% decodable to be considered decodable, which seems to align with some research (e.g., Jenkins and colleagues (2004) defined decodable texts as at least 79% decodable words). Thus, I used a threshold of 80% decodable or high frequency words in order to give children substantial support and opportunity to apply letter-sound knowledge, while reserving a proportion of words to ensure meaningfulness and include content vocabulary.

Lesson-To-Text-Match. Most texts (32 out of 40) were connected to a specific week in a phonics scope and sequence (in this case that of the phonics program, Fundations®, which is mostly arranged with weekly phonics foci). Each text contained multiple opportunities for students to practice using particular letter-sound relationships while reading, aiming for a high level of lesson-to-text-match (Mesmer, 2005) to allow children many opportunities to contextualize and use specific letter-sound knowledge. The first eight texts were matched to kindergarten phonics standards to support children practicing using basic alphabet knowledge to decode words and to allow teachers to differentially support children's needs.

High Frequency Words. In addition to including high frequency words in the above measure of decodability, I aimed to have a high proportion of high frequency words. Following empirical work on texts such as Compton and colleagues (2004), Mesmer (2010), and Hiebert and Fisher (2007) and theoretical writings from Mesmer and colleagues (2012), I included many high frequency words to support word reading fluency. In Appendix 1.B Table 1, I show the proportion of high frequency words in each text based on two systems: first, the high frequency words taught in the district phonics program; and second, the 100 most frequent words as described by Fry (1980). I chose this list as because it was originally generated based on word frequencies for all word types (and parts of speech) in texts written for children, includes inflected morphemes of words, and is available for free.

Word Repetitions. While maintaining natural (or close to natural) syntax and language, I aimed to repeat words across a text. I measured this repetition through type-token analysis (a measure used by others; see Cunningham et al., 2005), a measure of the number of unique words in a text divided by the total running words in a text. The mean type-token ratio of texts was 0.44. Ideally, this level of repetition, though not as high as some texts, hopefully allowed children multiple attempts to decode words across a unit to build orthographic maps while maintaining natural, meaningful language and syntax. I also avoided including non-decodable singlets when possible to decrease children's cognitive effort on non-decodable words (across the texts, 12.49% of singlets were not considered decodable words).

Word Understandability. Words deemed not decodable for a specific text were concrete, imageable, and familiar when possible (Fitzgerald et al., 2015) to support children's reading and understanding of these words. Words in each text that were not deemed decoded were, on average, highly familiar (95.66%), imageable (75.64%), and mostly concrete (68.39%).

In this case, words were deemed familiar if they were unit vocabulary words (e.g., *inventor, urban, compass*) or were words that the research team thought were likely to be recognizable to first graders in this district (e.g., *aquarium*, *litter*).

Text-Level Criteria

Conceptually Coherent Sets. Texts were organized in conceptual sets, based on the district's integrated literacy, science, and social studies curriculum (*FOF*). Each text matched to the conceptual curriculum through the topics and content vocabulary addressed. In this way, these texts attempted to support coherent knowledge building across a unit (Cervetti et al., 2016). In addition to including content vocabulary (such as *invent* and *create*), these texts directly linked to texts used for content read-alouds and genre features studied in writing. To support children in development of independent reading in the multitude of texts types and genres introduced in a particular unit in *FOF*, texts represented multiple texts types and genres (Duke & Roberts, 2010). For example, in one *FOF* unit focused on the weather and on persuasive letter writing, a Beyond Decodables text featured a fictional letter from a first-grade class in Boston, attempting to persuade the superintendent to give them a snow day.

Background, Cultural, and Place-Based Knowledge. All texts were written with consideration to children's backgrounds. Across the year, the texts depicted characters from a multiplicity of racial and ethnic groups, with high representation of the backgrounds most common among students in the district. In addition to helping children see themselves and their lives in texts, this may support reading fluency and comprehension (e.g., Bell & Clark, 1998; Priebe et al., 2012). Texts were also set in locations likely to be familiar to children in the district, such as one story set on a city bus, as relevant prior knowledge may contribute to increasing comprehension.

Enjoyable. All texts were written and revised to be enjoyable for students. For example, children during the pilot year (the year prior to this study year) enjoyed serial books, so I included several series of texts with the same characters. This may support students by encouraging students to persist through challenge (Fulmer & Frijters, 2011). Further, by creating texts meant to be engaging for students, these texts challenge the contention that texts with a high level of decodability cannot maintain children's interest in reading. Thus, by designing Beyond Decodables texts with attention to what is likely to be interesting for students, these texts may contribute to children's motivation to read; or, at least, may not detract from it as current decodable texts might (Castles et al., 2018).

Professional Development

Teachers received three major types (professional development, lesson planning materials, and text guides) of supports to implement these texts. All of these supports were available on a password-protected website, accessible throughout the year: lesson guides for each text and three types of lesson templates for each setting (whole group, small group, and independent/home reading logs).

Teachers received one one-hour professional development session (given by a district literacy leader and me) at the beginning of the school year to learn how to appropriately use these texts in small group reading lessons and support independent reading. Additional professional development and coaching was planned, but was not possible due to the COVID-19 pandemic

Implementation

Teachers were asked to use at least one decodable text per week with the majority of children in their classroom, primarily through small group instruction, beginning in the last week

of September. I recommended three lesson structures based on achievement: whole group shared reading if a text was too complex for a given class, 3-step small group reading for children who needed some support accessing the text (top recommendation), and a whole- or small-group launch for partner reading for children who were likely able to access the text on their own. Implementation is described further in the Outcomes section. The start date overlapped with the pre-test date; however, it is unlikely that the impact of the text would be due to one week. Within the first weeks of implementation, teachers received one in-person, professional development session about the texts (described above).

Sample

Due to the impacts of the COVID-19 pandemic, the study focuses on the implementation and impact of these texts during the first half of the school year (September 2019-December 2019). The school district provided administrative data and reading achievement data for first graders. The reading achievement data came from the districtwide early literacy assessment, NWEA MAP Reading Fluency, an assessment for K-2 early literacy skills.

Total Sample Schools

This study is a clustered observational study. The treatment was offered non-randomly to all schools in the district serving first-graders that fit the following requirements: using the district-created curriculum Focus on First, the phonics curriculum Fundations®, and NWEA MAP Reading Fluency. The total sample, consisting of the treatment schools ($N = 10$) and possible comparison schools ($N = 26$), is 36 schools.

Classrooms in treatment and comparisons schools in this analysis were regular education classrooms as I excluded small classrooms (less than 10 students) in which 100% of children had Individual Education Plans (IEPs). This led to 96 regular education classrooms.

First graders in the total sample took at least one NWEA MAP Reading Fluency Adaptive Oral Reading assessment including the phonics portion or oral reading fluency portions during the September (pretest) or December (posttest) testing periods. I defined these children as the sample to mitigate potential errors in district-level reporting (e.g., this allowed me to define the sample as children for whom there was a record of attending their school and class for at least one day, the testing day). This led to 1,604 children who represent multiple racial, ethnic, and linguistic groups (30.36% Black, 9.22% Asian, 39.15 % Hispanic, 16.08% White, and 5.17% Native American, Multiracial, and Other; 39.84% English Language Learners). Sample children tend to be classified as "economically disadvantaged" (76.12%), a district designation based on a student's participation in one or more of the following state-administered programs: the Supplemental Nutrition Assistance Program (SNAP); the Transitional Assistance for Families with Dependent Children (TAFDC); the Department of Children and Families' (DCF) foster care program; and MassHealth (Medicaid). See Table 1 for more demographic characteristics for the full sample.

Sample versus other district schools. There are 78 schools in the Boston Public Schools that have first-grade classrooms. In general, based on publicly available school-level data, the 36 sample schools in this study are statistically similar to all other Boston schools serving first-grade students. These similarities include: similar racial, ethnic, and linguistic demographics, total and first-grade enrollment, attendance, and, for those schools with available data, state ELA test scores. Out of these 78 schools total, 49 schools used the NWEA MAP Reading Fluency assessment. There was no statistically significant difference between schools in the sample and schools not in the sample (see Appendix 1.C).

Treatment versus comparison schools. Eleven schools volunteered to participate. One school did not begin the intervention until January 2020 (after the winter test date), so I dropped this school from analysis. One teacher opted out of participation and one volunteer teacher taught in a self-contained classroom, so I dropped these classrooms from analysis. Ultimately, 25 general education teachers at ten schools implemented at least one decodable text in the fall of 2019 (n = 425 children).

The ten treatment schools served first graders who were similar to children in the 26 comparison schools in age, gender composition, and the proportion of children with IEPs and ELL status (see Table 1 for descriptive statistics of treatment and comparison schools, with listwise deletion of missing data). Children had statistically similar pretest scores. First graders at treatment schools were statistically significantly more likely to be designated as economically disadvantaged. The racial and ethnic composition of first graders in participating and comparison schools was different; children in the treatment schools were statistically significantly more likely to be Hispanic and less likely to be Asian.

Procedures

The Institutional Review Board of the University of Michigan deemed this study exempt from oversight through the exemption for educational research involving normal educational practices**.**

Treatment School Recruitment

A district partner invited schools meeting the following criteria to participate: 1) serving first graders in the district to voluntarily participate in this study, 2) consistently and accurately using Northwest Evaluation Association Measure of Academic Progress (NWEA MAP) Reading

Fluency to measure student achievement in early literacy,[3] 3) using the district-created standardized curriculum *Focus on First* (*FOF*) in combination with Wilson Fundations®, a systematic phonics program, and 4) did not serve as a pilot school in the 2018-2019 school year ($N = 2$ schools). Thirty-seven schools out of 78 schools serving first graders in Boston were eligible to participate. Eleven schools volunteered to participate as treatment schools in the 2019-2020 school year (ten implemented the project). All teachers at volunteer schools were also asked to participate and consent to the research. Ninety-seven percent of treatment teachers consented to participating in this research.

Student Recruitment

Individual students and their families were not required to consent to research as this additional reading practice is not outside of typical educational practices. I only use administrative, deidentified data about children in this study.

Data Collection

Student Data. Children were assessed individually by the NWEA Map Reading Fluency Adaptive Oral Reading assessment (NWEA, 2019), a computer adaptive assessment and, if necessary, with support from their classroom teacher. This assessment was recommended, though not mandated, by the district regardless of study participation. Approximately 12% of children were assessed outside the district-defined pre-test date (September). In the main analysis, I include all children even those who took the pre-test outside of the testing data (see Appendix 1.D for a robustness check including only children who took the assessment in the

[3] District policy changed at the beginning of the 2019-2020 school year, so schools were no longer compelled to assess literacy in K-2. Thus, schools differed in their use of this assessment. Schools were only included in this sample if they assessed most children on record with the full NWEA MAP Reading Fluency Adaptative Oral Reading assessment during the fall/winter 2020.

correct time period). The district research office provided administrative, deidentified assessment scores and demographic information for al first graders.

Fidelity. To understand intervention fidelity (the extent to which the program was implemented as designed; Hulleman et al., 2013), teachers in the treatment group participated in a survey and observation.

Fidelity Survey. All teachers at treatment schools who consented to participate filled out an electronic survey in December 2020 (See Appendix 1.E Table 3). Teachers were emailed individual links to fill out the survey and sent two reminder emails. Eighty-four percent of teachers filled out the survey. Teachers (n =21) answered survey questions about their demographic characteristics (racial/ethnic identities, years of teaching experience, highest earned degree) and teaching with decodable texts.

In the survey, teachers selected the decodable texts their students had read over the year and then answered a series of questions about implementation within the last two instructional weeks. These questions include asking teachers to reflect on how many students interacted with a text and in what context, their perception of student engagement with texts, and an open-ended portion to leave comments or ask questions. Although research suggests that teacher ratings of fidelity are not always as accurate as other ratings (Domitrovich et al., 2010), teacher self-surveys yielded information about implementation that could not be captured in the observations.

Fidelity Observation. Teachers who consented to participate were scheduled to be observed once in December 2020 or January 2021, depending on individual scheduling constraints (See Appendix 1.E Table 2). Teachers filled out a survey listing their time preference for observations and were alerted prior to the observational visit. Prior to the observations, I created an intervention fidelity coding scheme with both quantitative and qualitative elements to

understand teacher's fidelity to intervention materials and teachers enactment of texts (to support greater district understanding about teaching with these materials). I trained two district early literacy staff members on the coding scheme. Then, we coded two videos of lessons and made minimal revisions to the coding scheme to better capture instruction until all three coders agreed on all scores for these two videos. We then observed two classrooms together, meeting between each observation to compare and discuss difference in coding. In the next five observations all three coders, we agreed 80.95% of the time. Achieving 80% agreement, we continued observing individually and in groups of two or three. Due to district research policies, teachers were able to view the fidelity coding scheme prior to their observation. We then observed and coded three teachers across two schools to test the fidelity tool and come to consensus. Twenty-one treatment teachers were observed by one to three trained coders (two district coaches and me) on an intervention fidelity coding scheme developed for this study. I averaged the score for each question on the fidelity tool when teachers were observed by more than one observer.

Composite Fidelity Measure. I then created a composite score of fidelity, adding together scores from the thirteen surveyed or observable actions that aligned with the resources provided to teachers and the research above. These thirteen actions centered on whether teachers were using the lesson templates and professional development suggestions when teaching with decodable texts (see Appendix 1.E Table 1 for the composite fidelity tool). For example, based on the above research, I encouraged teachers to introduce new high frequency words by analyzing the word's structure (rather than simply saying the word), so one question asked how teachers introduced new words (1 = with word's structure; 0 = through "sight" or did not introduce).

Measures

Outcomes

I used children's NWEA MAP Reading Fluency Adaptive Oral Reading (NWEA, 2019) assessment phonics scores and oral reading fluency scores as measures of literacy achievement. This assessment was already administered by all schools, making it an easy lift for teachers, free for the research budget, and an easy sell to district research evaluators. The post-treatment assessment was administered in December 2019. NWEA MAP Reading Fluency is an early literacy assessment that measures phonological awareness, phonics and word recognition, reading fluency, reading and oral language comprehension, vocabulary, and oral reading through a computer adaptive assessment (NWEA, 2019). Children either take a comprehensive foundational skills assessment or an oral reading fluency assessment (more advanced route) based on a separate screening tool.

Though this assessment assesses more than children's phonics and word reading knowledge, I chose to only use the phonics and oral reading fluency sub-score from the foundational skills assessment because a measure of word reading is the most conceptually related to the theoretical benefit of decodable texts. On this subscale, children are adaptively given 9-12 items per category, beginning with recognizing letters and sounds in isolation, recognizing and generating letters in words, reading and spelling regular consonant-vowel-consonant words, and reading and spelling regular one-syllable words. For example, a question might be "Which letter is *b*?" and the child is prompted to select between several letters visually. Another question might be, "Spell the word *cap*" and the child is prompted to use letter tiles to spell the word.

The phonics sub-score is a 1-4 scale based on the type of words children are able to recognize or spell, where 1 represents letters and sounds, 2 represents letters in words, 3

represents regular consonant-vowel-consonant words, and 4 represents regular one-syllable words. To compare children who took the oral reading fluency assessment on the same scale, I rated children who took the oral reading fluency assessment as 5 (the screener indicates their skill level is above the four phonics subscore categories; pretest *N*= 125; posttest *N* = 243). Internal evaluation suggests that NWEA MAP Reading Fluency is a reliable assessment of these literacy skills for children in grades K-3. NWEA MAP Reading Fluency also has demonstrated concurrent validity with NWEA MAP Reading Growth, a reliable and valid literacy assessment (NWEA, 2019). This assessment is new (it was released in 2018) and has not been used extensively in research.

Treatment indicator

I created a dummy indicator (treatment = 1) to indicate children participated in a treatment classroom (0 to indicate which children were in comparison classrooms). Teachers were given professional discretion to implement the texts in multiple settings and with different groups of children; therefore, treatment took place at the classroom level.

Covariates

I used children's phonics sub-scores on NWEA MAP Reading Fluency Adaptive Oral Reading from September as a control for children's pre-intervention skills. I used administrative dummy variables to indicate each child's race and ethnicity (White, Black, Hispanic, Asian, Other/Native American), English Language Learner status, disability status (1 = has an Individualized? Education Plan), gender (1 = male), and economic status (1 = economically disadvantaged) (see Table 2 for treatment and comparison children's demographic data). I also include a variable for the time between test periods (number of days), as this varied substantially (35-100 days) (all children in the treatment condition took the assessment during the correct

period; see Appendix 1.D for an additional robustness check with only children who took the pre-test prior to intervention implementation).

I constructed a classroom level variable for class size due to the large variation in size (5 to 25 children per class).

I constructed school-level data based on publicly available data from SY 2018-2019. I created to indicate the proportion of a school's population identified as male, White, Black, Hispanic, Asian, and Other/Native American/Mixed Race (the district determined the racial categories). I also create variables with each school's proportion of English Language Learners, students with individual education plans, and economically disadvantaged students (see Table 1 for more school demographic data).

For treatment schools, I collected additional classroom-level information. Teachers provided their race and ethnicity (all identified as Black, White, or Multi-racial), gender, and years of teaching experience (see Table 3 for teacher demographic data). These data are used for descriptive purposes only.

Fidelity of implementation

I investigated fidelity in three ways. The key fidelity measure was constructed from the fidelity observation and fidelity survey described earlier. Key indicators are shown in Table 4. First, I investigated higher versus lower fidelity of implementation. To construct two fidelity groups, I coded a binary indicator from the survey and observational data. I coded a teacher as having higher implementation (1 = higher implementation) if they met the following: 1) reported using intervention materials in the last 2 instructional weeks (demonstrating some consistent use); 2) had used more than 10 decodable texts in the first half of the school year (mean usage = 10; demonstrating higher-than-average usage); and 3) scored over 6.45 out of 15 on the

implementation rubric (mean score = 6.45, demonstrating observed quality above average). Second, I investigated fidelity on a continuous scale, using the implementation rubric scale. Third, due to the theoretical benefit of decodable texts when directly linked to phonics instruction, I investigated differences in outcomes based on if teachers scored "yes" on the observation question "The focus of the lesson is on using phonics knowledge while reading." During reliability rounds on the fidelity measure, we determined this question could only be answered "yes" if teachers explicitly supported children in using phonics knowledge and focused on phonics knowledge multiple times during the lesson. Four teachers scored "yes" on this question.

Analytic approach

RQ 1

To address RQ1 (*Do multiple-criteria, content-connected texts impact first-graders' word reading, as measured by NWEA Map Reading Fluency?*), I fit a multilevel model without matching in order to estimate the impact of the treatment with all possible treatment and comparison students. To do so, I first investigated the intraclass correlations (ICCs) for the winter word reading score. The ICC for schools was 5.98%. This indicates a moderate amount of variance between schools – variance that I adjust for using a two-level model with random intercepts at the school level. My primary model for doing so was:

$$Word\ Reading_{ijk} = \beta_0 + \beta_1 Treat_k + \beta_2 Pretest_{ijk} + \delta X_{ijk} + \gamma ClassSize_{jk} + \rho School_k + (\nu_k + \varepsilon_{ijk}) \quad (1),$$

where the subscripts *i*, *j*, and *k* refer to students, classrooms, and schools respectively; *Word Reading* is the child's phonics score in the winter of first-grade from the administrative NWEA

data; *Treat* is a school-level dichotomous variable set to 1 if the school participated in the intervention and 0 otherwise; *Pretest* is the child-level pretest phonics score in the fall of first grade; *X* is a vector of child-level characteristics (age, gender, race/ethnicity, IEP status, ELL status, economically disadvantaged status, and time in days between test dates); *ClassSize* is a measure of class size; *School* is a vector of school-level demographic covariates (proportions of students attending a school based on race/ethnicity, gender, IEP, ELL, and economically disadvantaged); ν is a school-level random intercept; and ε is a class-level and student-level random error term. β_1 is the estimate of the impact of the decodable text intervention on each outcome.

I next used a more sophisticated, new multi-level matching approach following Keele and colleagues (2020) and Page and colleagues (2020). Despite control variables included in equation (1), my estimate of the effect of the treatment may be biased because selection into treatment may have been driven by some unobserved factors, such as a principal's orientation towards research. Though cluster-level assignment may reduce some potential bias (Hansen et al., 2014), selection bias is a major threat to the internal validity of findings of clustered observational studies.

To attenuate bias in clustered observational studies, researchers often turn to regression adjustment or propensity score matching. As outlined by Page and colleagues (2020), both strategies have drawbacks. Regression adjustment, on its own, may violate the assumption of "common support," which assumes overlap in the chance all schools had of a chance receiving treatment. Propensity score matching often fails to converge or does not appropriately fit with multilevel data. Page and colleagues (2020), along with others (e.g., Keele et al., 2020; Pimentel et al., 2018; Zubizarreta & Keele, 2017) suggest a new matching algorithm that takes into

account the multilevel structure of the data and decreases the probability of violating assumptions necessary for the validity of interpreting clustered observational studies.

Accordingly, following Keele and colleagues (2020) and Page and colleagues (2020), I used a multilevel matching algorithm to match treatment schools and students to comparison schools. In this approach, the algorithm first matches students at treatment and comparison schools based on student-level covariates, much like typical matching procedures. Then, the program iteratively computes school-level matches based on decreases the distance between student-level matches and specified school-level covariates and produces an optimal match. I matched comparison and treatment schools on all above student-level covariates in equation (1). To improve balance further, I then specified "fine balance" on three school-level covariates, the proportion of students who are identified as male, Asian, and Hispanic, that were particularly poorly balanced, even after matching at the student-level. "Fine balance" ensures that the distributions in each category between treatment and control units are similar. As recommended by Page and colleagues (2020), I transformed the three covariates that I used for refined covariate balance to binary variables, cut at the mean (where 1 = higher than mean proportion of the student population).

In order to allow for optimal matches (excludes students who are more difficult to match, defined as further than 0.05 quantile away), it was also necessary to trim some student observations from both the comparison and treatment groups. This ensured that there were equal numbers of students in the treatment and comparison groups. Therefore, estimations using multilevel matching represent the estimated treatment effect on a subset of approximately 70% of the students in the treatment group only. Though I did not achieve optimal balance on all observables (optimal is often thought of as standardized differences of less than 0.10; Pimentel et

al., 2018), balance was greatly improved on school-level covariates through this fine matching (see Table 2).

After trimming, I used the multilevel matching technique described above in combination with the multilevel regression model (equation 1). When using multilevel matching, I dropped student-level covariates (except the pretest) and the fine balanced school-level covariates from the model, as these were already accounted for in the matching technique. Further, when including the multilevel matches in the multilevel regression model, I also included a random intercept for matched pairs to account for the newly created nesting of schools within matched pairs (as suggested by Page et al., 2020).

RQ 2

To answer research question 2 (*Is higher fidelity to implementation positively associated with gains in first-graders' word reading?*), I first compared the "higher" and "lower" fidelity classrooms on each item on the survey and observation tools and report standardized differences and significance levels based on t-tests. I then used a residual gains approach. I used equation 1 (above), but only included treatment classrooms. Instead of a school-level dichotomous variable set to 1 if the school participated in the intervention and 0 otherwise, I included a dichotomous variable for implementation quality set to 1 if the teacher implemented the intervention with higher fidelity to implementation (define above). Next, I compared on the continuous observation fidelity scale using the same model as above (following Duke et al., 2021). Finally, I compared outcomes in classrooms where teachers' instruction focused on using phonics while reading (coded 1) to other treatment classrooms (coded 0). I used the same equation described above.

Missing Data

Partially due to the constraints on the data (such as including only children who had demographic information and at least one recorded test) included in this analysis and the use of administrative data, the rates of missingness were relatively low, with 4.99% missingness on the pretest and 8.29% missingness on the outcome. There was no missingness on any covariates.

The 213 children with missing data were different than those without. Children with missing data were statistically significantly more likely to be male and English Language Learners (see Table 2 for missing data). Despite the differences in children with and without missing data, given the relatively low rates of missingness and attrition, I use complete case analysis. After listwise deletion, the final sample was 1,391 children in 93 classrooms in 36 schools. In Appendix 1.D, I present a robustness check using the dummy variable method for the predictor and outcome, substituting the class mean on both tests to account for missingness.

I also investigated attrition to ensure rates did not dramatically differ in the treatment and comparison groups. There was no classroom- or school- level attrition (beyond the volunteer school discussed above that did not implement the treatment). At the child-level, 149 children did not complete the post-test (8.21%). In the treatment group, 5.17% attrited and in the comparison group, 9.65% attrited. This level of overall attrition and differential attrition rates are acceptable (WWC, 2017).

Sensitivity Analyses

As previously referenced, I used a variety of mechanisms for checking the sensitivity of my results. All robustness checks are presented in Appendix 1.D. First, in order to increase the possibility of balanced matches, I drew a comparison group sample from all non-treatment schools in the district using the same reading assessment, regardless of other criteria. Second, in order to avoid biasing the treatment effect due to some teachers administering the pre-test after

the official district testing period (children were assessed between early September and late October), I present models only including children in treatment and comparison schools whose pretest was administered before the intervention began. Third, I present a model using the dummy variable method for children's test scores to differently account for missingness on the pre-test, as recommended by Puma and colleagues (2009). This method is less likely to bias the estimates and standard errors than other methods. I constructed a variable set to one for children with missing data and set the value of the missing pretest score to zero. This allowed me to use more of the available records in this sensitivity check. Fourth, to account for children's nesting within classrooms, I investigated the intraclass correlations (ICCs) for the winter word reading score in a three-level model. The intraclass correlation for a three-level model for classrooms was 11.45% and for schools was 1.90%, so I present a three-level model with random intercepts for both schools and classrooms.

Results

Matching and Balance Check

Table 2 provides balance checks with standard differences for baseline measures for the three main models for research question 1 (including all comparisons, matching by student-level covariates, matching with school-level refined balance). These checks show the improved balance between treatment and comparison classrooms with the multilevel matching procedure. In the final match with refined covariate balance, there are no statistically or marginally significant differences between any observed covariates, though the magnitude of some differences is above 0.10, the recommended ideal balance (Page et al., 2020).

Research Question 1

Table 3 summarizes the estimated impact of the intervention on children's phonics outcomes, across the three models. The first model shows the estimated impact of the invention compared to all other schools, without matching (95% CI [-0.12, 0.09]; β_1 = -0.01). The second model shows the estimated impact of the intervention with matching based on student-level covariates (95% CI [-1.61, 1.43]). This model was less precise than the first, though with a similar estimate (β_1 = -0.01). The third model shows the estimated impact of the intervention with matching at the student-level and fine-balance on school-level covariates (95% CI [-0.26, 0.34]). The third model was more precise than the second. The third model's treatment effect was positive ((β_1 = 0.04); however, it was not substantially different from the first two models. The treatment estimates were relatively stable, as were effect sizes (as follows: 0.05 for model 1, 0.03 for model 2 and 0.08 for model 3).

Research Question 2

Table 4 summarizes the results of the fidelity survey and observational rubric. Fidelity varied widely, both in observations and self-reports. Teachers who responded to the survey reported implementing between one and 17 texts in the fall (*SD*=3.86) and using them in the two weeks prior to the survey with no children (4.76%), some children (38.10%), or all children (57.00%). Observed teachers varied in their practices as well, scoring between 3 and 12 on the implementation rubric (where the highest possible score would be 15).

Implementation as Binary Indicator

Teachers in the "higher quality implementation" category were statistically significantly more likely to introduce new words by examining a word's structure with students and prompt students with reminders of their phonics knowledge. They were also marginally more likely to prompt students to use word-reading strategies such as "tap it out" or "say each sound." Further,

children in the classrooms of teachers with higher implementation tended to be more likely to be observed using phonics knowledge to read words. None of these observed differences were highly or moderately correlated with changes in phonics scores (individual fidelity items were correlated with changes in phonics score from 0.01-0.08).

Table 5 Model 1 summarizes the association of higher-quality implementation of the intervention with children's phonics outcomes. There were no statistically significant associations of high-quality implementation of the intervention at the middle of first grade compared to lower quality implementation (95% CI [-0.25, 0.27]; β_1 = 0.01). This estimate is similar to those in Table 3, further indicating a lack of association of word reading outcomes to observed and surveyed fidelity.

Implementation as Continuous Indicator

Table 5 Model 2 summarizes the association of the continuous observation implementation scale score with children's phonics outcomes. There were no statistically significant associations of the scale with phonics outcomes (95% CI [-0.25, 0.27]; β_1 = 0.01). This estimate is similar to the results of the binary fidelity indicator and those in Table 3, further indicating a lack of association of word reading outcomes to observed fidelity.

Implementation as Indicated by Focus on Phonics

Teachers who scored "yes" on the question "the focus of the lesson is on using phonics knowledge while reading" taught in classrooms that were different than other observed treatment classrooms (*N* =22). These students were statistically significantly less economically disadvantaged (70.27% versus 82.17%) and had statistically significantly higher pre-test phonics scores (2.99 versus 2.74). Classrooms were teachers scored "yes" were also had statistically significantly larger than other classrooms (19.80 children versus 16.67 children). In other words,

these classrooms cannot be considered accurate counterfactuals of one another. Despite these differences and other unobserved confounding factors, I did compare these outcomes due to the theoretical interest of the implementation question ("the focus of the lesson is on using phonics knowledge while reading").

Table 5 Model 3 summarizes the association of focuses on phonics while reading and with children's phonics outcomes. There was a marginally statistically significant association of focuses on phonics compared to not focusing on phonics in observations (95% CI [-0.02, 0.49]; $\beta_1 = 0.23$).

Sensitivity Analyses

The sensitivity analyses/robustness checks are described in Appendix 1.D. Results were nearly identical with different comparison groups. Results were robust to changes in assumptions and modeling.

Discussion

This study attempted to give additional policy-relevant context to the on-going debate about using decodable texts. Due to COVID-19 disruptions, the results of this quasi-experimental study do not provide evidence for or against the continued use of decodable texts. Though results of this study do not provide clarity on about the impact of reading decodable or multiple criteria texts in first grade, this study contributes to the growing literature on decodable texts in several critical ways.

First, this study was the first to attempt to study a "light-touch" implementation of decodable or multiple criteria texts across a large district. Though the present study could not examine the impact of that implementation over the course of a year, as intended, it did point to the possibility of additional longer, large-scale studies of decodable text. With some additional

professional development, other adjustments, and without the impacts of a global pandemic, a future study may find positive impacts or may find no or negative impacts—in any case useful information for the field.

Second, this study was the first to use these particular factors (e.g., 80% or more decodable words, connected to content, supportive of culturally responsive instruction) in creating texts. Though these are perhaps not the only or most optimal text scaffolds to support developing readers, this study demonstrates that multiple criteria texts can (or at least can attempt to) support a knowledge-building, culturally responsive curriculum. Responding to criticism about decodable texts lacking content (such as Castles et al., 2018; Martinez & McGee, 2000), this study, along with others on multiple criteria texts (Cheatham et al., 2014), demonstrate that it is possible to write texts that are highly decodable, meaningful, and place-based.

Third, though exploratory, preliminary evidence suggests that teachers may be best able to support children's word reading development by using decodable and multiple criteria texts with in-the-moment, explicit phonics supports. This extends prior theoretical work that suggests decodable texts complement systematic, explicit phonics instruction by indicating that children may need additional supports or reminders to use this knowledge in text reading. In future studies, researchers should aim to further investigate how differences in implementation and instruction with multiple criteria texts impact word reading outcomes.

Finally, this study makes a methodological contribution in demonstrating the potential utility of using a procedure for multilevel matching to improve the balance between treatment and comparison groups in clustered observational studies. Though even with improved balance, this did not yield definitive results, this procedure, when adding additional fine balance at the

school level, led to greater precision (Table 3, Model 3) than matching on student-level covariates. With larger samples and additional observed covariates, researchers may be able to approximate a randomized control trial more closely.

Returning to the earliest study on decodable texts reviewed in this article, Juel and Roper-Schneider noted a major issue in reading instruction in 1985 that still persists today, "Children may be instructed to recognize words by "sounding them out," using letter-sound correspondence taught in phonics. Yet, they may read from a basal text with few regular decodable words. That is, they may see many words which neither respond well to the strategy they are learning, nor provide practice in reading words with similar sound patterns" (p. 136). In elementary classrooms in this country, it may still be the case that children are taught phonics but lack the opportunity to use phonics in texts.

Decodable texts reflect an attempt to answer this need, and, later, in response to the need to create high-quality texts that support beginning readers across many aspects of reading, multiple-criteria texts attempted to fix this issue. Much like the research on decodable texts more broadly, the evidence that multiple-criteria texts are superior to other texts is preliminary, though it does indicate some promise of these texts for beginning readers. This study found no clear impact of multiple-criteria, content-connected text reading on first grade word reading outcomes, but did find early exploratory evidence that phonics-focused instruction along with multiple criteria texts may support word reading outcomes.

Limitations

First, due to the impacts of COVID-19, this study did not proceed as intended. Though teachers used Decodables 2.0 through mid-March and received additional individualized professional development and coaching in February, the timing of the district assessment meant

that the impact of the intervention could only be studied from October to December 2019 (see Figure 1 timeline). Rather than studying the impact of a year-long intervention, I was only able to investigate gains children made in an average of 11 weeks, interacting with texts for about 10 minutes per week, so it is unsurprising that an intervention of, at best, approximately 110 minutes total, with minimal support for teachers, did not led to substantial gains for children.

Second, the NWEA MAP Reading Fluency assessment was not particularly sensitive to changes in children's word reading. Most notably, in this sample, 45% of children demonstrated no change from September to December on the measures of word reading. Though this measure may be valid and reliable for measuring longer-term growth, this measure did not measure growth in this particular sample of children in this brief time period. As it is unlikely that nearly half of children in a district did not improve their word reading in the first half of first grade, one major reason for this null impact may be the use of this blunt instrument. In addition, in this sample, only 14.5% took the oral reading fluency portion of the assessment, which may be a more potent measure of growth as the oral reading fluency portion contains continuous scores for words correct per minute and accuracy (instead of reporting levels, as the phonics portion does). Future research should use caution when using this measure to assess small instructional changes.

Further, the fidelity tools may not have been sensitive enough to pick up differences in implementation. Across teachers, there was minimal variation in all survey questions (with the exception of the number of texts read), attention to and connections to content learning, and giving children access to texts. Though there were differences in how teachers instructed children to read and prompted students during reading compared to the intended instruction, the

observation tool (and limitations of only having one observation) may not have illuminated the most critical differences between teachers.

Third, the attempt to add a "light-touch" implementation may simply not support teachers' needs in implementing a new type of text and instruction with high fidelity and quality. There are several reasons why simply adding some multiple criteria texts to a teacher's repertoire of instruction moves may not impact children's outcomes. First, Juel and Roper-Schneider (1985) showed that the initial method of instruction and text type impacted the word recognition strategies that children used for the entire first grade year. In other words, instruction through a different schema than is ideal for decodable texts may have led children to use guessing, picture clues, and/or rely on memorization, even when reading decodable texts. Though teachers with higher-quality implementation scores were marginally statistically significantly more likely to have children use phonics-based strategies and skills while reading decodable texts in the observation (standardized difference = 0.81; see Table 4), children may not have generalized this skill into other reading contexts. In one instance, for example, a teacher (whose other scores meant her instruction fell into the higher-quality implementation group) prompted children to "guess" a word based on the picture even though children in the group had already successfully decoded the word, which potentially sent children the message that using the pictures was a more important skill than decoding. Without additional support for teachers to pivot instruction, introducing a light-touch addition of multiple criteria texts may not have improved children's use of phonics-based strategies and skills enough to influence reading outcomes.

Finally, this study is limited by its design. Schools who opted into treatment are likely fundamentally different than comparison schools by unobservable characteristics. Indeed, even when attempting to control for variation with propensity score match and weighting, differences

remained between the two groups in observable characteristics. Without randomization, it is difficult to estimate the impact on children in treatment schools.

Future research

Research is necessary to clarify the impact of decodable and multiple criteria texts. At present, though the theoretical evidence is strong, there is simply not enough evidence to conclude that these texts are more supportive of word reading development than other texts. Critically, however, there is not sufficient evidence that they are equally or less supportive either. The search for the optimal scaffolds within text continues.

In a year unencumbered by a global health crisis and with additional funding and with a district commitment to randomization, a researcher could replicate and extend this study to understand the potential impact of content-connected, multiple criteria texts more precisely. First, ideally, a study would use randomization in order to compare impacts without the need for matching techniques. Second, in order to address policy-level implications, a new study would need to recruit far more schools and teachers than this study. Third, to more carefully study implementation, surveys and observations should estimate the ratio of multiple criteria versus other texts read by individual students, describe with more detail the control condition (in particular, whether any decodable texts were used), and examine more differences in teacher instruction, prompting, and language while using the texts. A resource-intensive study could even use procedures such as the Individualizing Student Instruction (Connor et al., 2009) to examine the impact of individual students' dosage of multiple criteria texts on outcomes.

One essential component of this body of research that needs clarity is the definition of a decodable word, which is not consistent across studies; some research defines decodable words as graphophonemically regular words, regardless of a child's knowledge or instructional context,

whereas other research defines decodable words as words with only graphophonemic patterns known to the reader at that point in time (e.g., definitions discussed by Murray et al., 2014). Although both make assumptions about children's knowledge, the former is not directly related to children's learning and therefore may be less likely to be directly supportive of children's application of phonics in context. Future research should seek to clarify how much decodability is necessary.

A second essential component of this body of research in need of clarity is the variation in teaching practices. Fundamentally, these studies all ask, "Is reading decodable texts along with an explicit, systematic phonics program supportive of reading development?" Most studies, including this one, do not report what else teachers are doing throughout the school day as they teach reading. Future research should focus more on the instructional context surrounding the texts and teachers' understandings of decodable texts. Ideally, future research should attempt to intervene with respect to children's entire experiences around word reading (phonics, small group reading, texts, teacher prompts, etc.) to investigate whether theoretical consistency across reading instruction, rather than an intervention regarding texts alone, could, in fact, lead to better reading outcomes.

References

Amendum, S. J., Conradi, K., & Hiebert, E. (2018). Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension. *Educational Psychology Review*, *30*(1), 121-151

Allor, J. H., Yovanoff, P., Al Otaiba, S., Ortiz, M. B., & Conner, C. (2020). Evidence for a Literacy Intervention for Students with Intellectual and Developmental Disabilities. *Education and Training in Autism and Developmental Disabilities*, *55*(3), 290-302.

Bell, Y. R., & Clark, T. R. (1998). Culturally relevant reading material as related to comprehension and recall in African American children. *Journal of Black Psychology*, *24*(4), 455-475.

Beverly, B. L., Giles, R. M., & Buck, K. L. (2009). First-grade reading gains following enrichment: Phonics plus decodable texts compared to authentic literature read aloud. *Reading Improvement*, *46*(4), 191-206.

Bowey, J. A., & Muller, D. (2005). Phonological recoding and rapid orthographic learning in third-graders' silent reading: A critical test of the self-teaching hypothesis. *Journal of experimental child psychology*, *92*(3), 203-219.

Caravolas, M., Lervåg, A., Mikulajová, M., Defior, S., Seidlová-Málková, G., & Hulme, C. (2019). A cross-linguistic, longitudinal study of the foundations of decoding and reading comprehension ability. *Scientific Studies of Reading*, *23*(5), 386-402.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51.

Cervetti, G. N., & Hiebert, E. H. (2019). Knowledge at the center of English language arts instruction. *The Reading Teacher*, *72*(4), 499-507.

Cervetti, G. N., Wright, T. S., & Hwang, H. (2016). Conceptual coherence, comprehension, and vocabulary acquisition: A knowledge effect?. *Reading and Writing*, *29*(4), 761-779.

Cheatham, J., Allor, J., & Roberts, J. (2014). How does independent practice of multiple-criteria text influence the reading performance and development of second graders? *Learning Disability Quarterly, 37*(1), 3-14.

Chu, M. C., & Chen, S. H. (2014). Comparison of the effects of two phonics training programs on L2 word reading. *Psychological Reports*, *114*(1), 272-291.

Compton, D. L., Appleton, A. C., & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor readers. *Learning Disabilities Research and Practice, 19*, 176–184.

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., ... & Schatschneider, C. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher*, *38*(2), 85-99.

Cunningham, P.M. & Cunningham, J.W. (2002). What we know about how to teach phonics. In A.E. Farstrup & S.J. Samuels (Eds*.), What Research Has to Say About Reading Instruction* (3rd ed., pp. 87–109).

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33,* 934-945.

de Graaff, S., Bosman, A. M., Hasselman, F., & Verhoeven, L. (2009). Benefits of systematic phonics instruction. *Scientific Studies of Reading*, *13*(4), 318-333.

Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & DeRousie, R. M. S. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly*, *25*(3), 284-298.

Duke, N. K., & Roberts, K. L. (2010). The genre-specific nature of reading comprehension. In *The Routledge International Handbook of English, Language and Literacy teaching* (pp. 98-110). Routledge.

Duke, N. K., Halvorsen, A. L., Strachan, S. L., Kim, J., & Konstantopoulos, S. (2021). Putting PjBL to the test: The impact of project-based learning on second graders' social studies and literacy learning and motivation in low-SES school settings. *American Educational Research Journal*, *58*(1), 160-200.

EdWeek (2020). *Early reading instruction: Results of a national survey.* Retrieved from https://epe.brightspotcdn.com/32/4f/f63866df760fb20af52754fd07ff/ed-week-reading-instruction-survey-report-final-1-24-20.pdf

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of reading*, *9*(2), 167-188.

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading*, *18*(1), 5-21.

Ehri, L. C. (2020). The science of learning to read words: A case for systematic phonics instruction. *Reading Research Quarterly*, *55*, S45-S60.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology*, *107*(1), 4-28.

Foorman, B. R., Francis, D. J., Davidson, K. C., Harm, M. W., & Griffin, J. (2004). Variability in text features in six grade 1 basal reading programs. *Scientific Studies of Reading*, *8*(2), 167-197.

Fry, E. (1980). The new instant word list. *The Reading Teacher*, *34*(3), 284-289.

Fulmer, S. M., & Frijters, J. C. (2011). Motivation during an excessively challenging reading task: The buffering role of relative topic interest. *The Journal of Experimental Education*, *79*(2), 185-208.

Gourley, J. W. (1984). Discourse structure: Expectations of beginning readers and readability of text. *Journal of Reading Behavior*, *16*(3), 169-188.

Grainger, J. (2018). Orthographic processing: A 'mid-level' vision of reading: The 44th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, *71*(2), 335-359.

Hargis, C. H., & Gickling, E. E. (1978). The function of imagery in word recognition development. *The Reading Teacher*, *31*(8), 870-874.

Henbest, V. S., & Apel, K. (2017). Effective word reading instruction: What does the evidence tell us? *Communication Disorders Quarterly*, *39*(1), 303-311.

Hiebert, E. H., & Fisher, C. W. (2007). Critical word factor in texts for beginning readers. *The Journal of Educational Research*, *101*(1), 3-11.

Hiebert, E.H. (2017). The texts of literacy instruction: Obstacles to or opportunities for educational equity? *Literacy Research: Theory, Method, and Practice, 66*(1), 117–134.

Hiebert, E. H., & Fisher, C. W. (2007). Critical word factor in texts for beginning readers. *The Journal of Educational Research*, *101*(1), 3-11.

Hiebert, E., & Fisher, C. (2016). A comparison of the effects of two phonetically regular text types on young English learners' literacy. *Reading Research Report*. Retrieved from textproject.org

Hiebert, E. H., & Mesmer, H. A. E. (2013). Upping the ante of text complexity in the Common Core State Standards: Examining its potential impact on young readers. *Educational Researcher*, *42*(1), 44-51.

Hiebert, E. H., Martin, L. A., & Menon, S. (2005). Are there alternatives in reading textbooks? An examination of three beginning reading programs. *Reading & Writing Quarterly*, *21*(1), 7-32.

Hulleman, C. S., Rimm-Kaufman, & Abry, T. (2013). Innovative methodologies to explore implementation: Whole-part-whole – construct validity, measurement, and analytical issues for intervention fidelity assessment in education research. In T. Halle, A.Metz, & I.Martinez Beck (Eds.), *Applying Implementation Science in Early Childhood Programs and Systems* (pp. 65-93). Baltimore, MD: Brookes Publishing.

Hulme, C., & Snowling, M. J. (2013). Learning to read: What we know and what we need to understand better. *Child DevelopmentPperspectives*, *7*(1), 1-5.

Hwang, H., & Duke, N. K. (2020). Content counts and motivation matters: Reading comprehension in third-grade students who are English learners. *AERA Open*, *6*(1).

Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading*, *8*(1), 53-85.

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78,* 243–255.

Juel, C., & Roper-Schneider, D. (1985). The influence of basal readers on first grade reading. *Reading Research Quarterly*, 134-152.

Keele, L., Lenard, M., & Page, L. (2021). Matching methods for clustered observational studies in education. *Journal of Research on Educational Effectiveness*, 1-30.

Lepola, J., Lynch, J., Kiuru, N., Laakkonen, E., & Niemi, P. (2016). Early oral language comprehension, task orientation, and foundational reading skills as predictors of grade 3 reading comprehension. *Reading Research Quarterly*, *51*(4), 373-390.

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *13*(2), 123-146.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247-288.

Menon, S., & Hiebert, E. H. (2005). A comparison of first graders' reading with little books or literature-based basal anthologies. *Reading Research Quarterly*, *40*(1), 12-38.

Mesmer, H. A. E. (2005). Text decodability and the first-grade reader. *Reading and Writing Quarterly, 21*, 61-86.

Mesmer, H. A. E. (2006). Beginning reading materials: A national survey of primary teachers' reported uses and beliefs. *Journal of Literacy Research*, *38*(4), 389-425.

Mesmer, H. A. E. (2009). Textual scaffolds for developing fluency in beginning readers: Accuracy and reading rate in qualitatively leveled and decodable text. *Literacy Research and Instruction*, *49*(1), 20-39.

Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, *47*(3), 235-258.

Murray, M. S., Munger, K. A., & Hiebert, E. H. (2014). An analysis of two reading intervention programs: How do the words, texts, and programs compare?. *The Elementary School Journal*, *114*(4), 479-500.

Nation, K., Angell, P., & Castles, A. (2007). Orthographic learning via self-teaching in children learning to read English: Effects of exposure, durability, and context. *Journal of Experimental Child Psychology*, *96*(1), 71-84.

NWEA. (2019). *MAP Reading Fluency Technical Report*. Portland, OR: NWEA.

Page, L. C., Lenard, M. A., & Keele, L. (2020). The Design of Clustered Observational Studies in Education. *AERA Open*, *6*(3). 1-14.

Pimentel, S. D., Page, L. C., Lenard, M., & Keele, L. (2018). Optimal multilevel matching using network flows: An application to a summer reading intervention. *Annals of Applied Statistics*, *12*(3), 1479-1505.

Pitcher, B., & Fang, Z. (2007). Can we trust levelled texts? An examination of their reliability and quality from a linguistic perspective. *Literacy*, *41*(1), 43-51.

Price-Mohr, R. M., & Price, C. B. (2018). Synthetic phonics and decodable instructional reading texts: How far do these support poor readers?. *Dyslexia*, *24*(2), 190-196.

Priebe, S. J., Keenan, J. M., & Miller, A. C. (2012). How prior knowledge affects word identification and comprehension. *Reading and Writing*, *25*(1), 131-149.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*, 546–567.

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, *46*(3), 343-366.

Rupley, W. H., Blair, T. R., & Nichols, W. D. (2009). Effective reading instruction for struggling readers: The role of direct/explicit teaching. *Reading & Writing Quarterly*, *25*(2-3), 125-138.

Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading*, *3*(3), 257-279.

Schiefele, U., Stutz, F., & Schaffner, E. (2016). Longitudinal relations between reading motivation and reading comprehension in the early elementary grades. *Learning and Individual Differences*, *51*, 49-58.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition, 55*, 151–218.

Simba Information (2017). *K-12 Reading market survey report 2017*. Simba Information.

Sparks, R. L., Patton, J., & Murdoch, A. (2014). Early reading success and its relationship to reading achievement and reading volume: Replication of '10 years later'. *Reading and Writing*, *27*(1), 189-211.

Stein, M., Johnson, B., & Gutlohn, L. (1999). Analyzing beginning reading programs: The relationship between decoding instruction and text. *Remedial and Special Education*, *20*(5), 275-287.

Taylor, B. M., Pearson, P. D., Clark, K., & Walpole, S. (2000). Effective schools and accomplished teachers: Lessons about primary-grade reading instruction in low-income schools. *The Elementary School Journal*, *101*(2), 121-165.

Torgerson, C., Brooks, G., Gascoine, L., & Higgins, S. (2018). Phonics: reading policy and the evidence of effectiveness from a systematic 'tertiary' review. *Research Papers in Education*, 1-31.

Torppa, M., Georgiou, G. K., Lerkkanen, M. K., Niemi, P., Poikkeus, A. M., & Nurmi, J. E. (2016). Examining the simple view of reading in a transparent orthography: A longitudinal study from kindergarten to grade 3. *Merrill-Palmer Quarterly*, *62*(2), 179-206

Tortorelli, L. S. (2019). Beyond first grade: examining word, sentence, and discourse text factors associated with oral reading rate in informational text in second grade. *Reading and Writing*, 1-28.

Williams, R., & Morris, R. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, *16*(1-2), 312-339.

What Works Clearinghouse (2018). *Reporting Guide for Study Authors: Group Design Studies*. Retrieved from https://ies.ed.gov/ncee/wwc/ReportingGuide?id=19

Zubizarreta, J. R., & Keele, L. (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *Journal of the American Statistical Association*, *112*(518), 547-560.

Table 2.1

Demographic Data by Treatment Status (SD), After Listwise Deletion

	Treat *n* = 381	Comparison *n* = 1,010	Standardized Difference
School Level			
Male %	0.50	0.53	-0.82*
	(0.02)	(0.03)	
IEP %	0.20	0.22	-0.37
	(0.06)	(0.06)	
ELL %	0.25	0.37	-0.21
	(0.12)	(0.13)	
Economically disadvantaged %	0.62	0.61	0.06
	(0.17)	(0.15)	
Race and ethnicity categories			
Black %	0.34	0.32	0.13*
	(0.19)	(0.16)	
Asian %	0.03	0.10	-0.75†
	(0.04)	(0.12)	
Hispanic %	0.43	0.37	0.37
	(0.16)	(0.16)	
Other %	0.05	0.05	0.00
	(0.02)	(0.02)	
Classroom Level			
Class size	17.53	17.72	-0.06
	(2.73)	(3.26)	
Child Level			
Age	6.15	6.13	0.06
	(0.35)	(0.34)	
Male	0.50	0.50	-0.01
	(0.50)	(0.50)	
IEP status	0.12	0.14	-0.05
	(0.33)	(0.35)	
ELL status	0.40	0.38	0.03
	(0.49)	(0.49)	
Economically disadvantaged status	0.82	0.74	0.19***
	(0.39)	(0.44)	
Race and ethnicity categories			
Black	0.32	0.31	0.03
	(0.47)	(0.46)	
Asian	0.04	0.11	-0.23***
	(0.21)	(0.31)	
Hispanic	0.45	0.36	0.18***
	(0.50)	(0.48)	
Other	0.04	0.05	-0.04
	(0.20)	(0.22)	

Days between tests	82.60 (5.53)	83.22 (9.81)	-0.08
Phonics pretest	2.72 (1.06)	2.66 (1.15)	0.06

Note. Economically disadvantaged status and the race and ethnicity categories are designations used by the district. T-test statistical significance levels to test for differences between groups are indicated as ***$p<.001$, **$p<.01$, *$p<.05$. Marginal statistical significance level indicated as $^{\dagger}p<.10$.

Table 2.2

Standardized Difference Between Treatment and Comparison Groups: Before Matching, Matching with Child-level Covariates, and Matching with Fine Balance

	No matching	Match with child-level covariates	Match with fine balance
School Level			
Male %	-0.82*	-0.91†	-0.13
IEP %	-0.37	-0.51	-0.16
ELL %	-0.21	-0.43	-0.07
Economically disadvantaged %	0.06	-0.55	-0.02
Race and ethnicity categories			
Black %	0.13*	-0.31	0.06
Asian %	-0.75†	-0.20	-0.16
Hispanic %	0.37	0.08	0.07
Other %	0.00	0.37	-0.13
Child Level			
Age	0.06	0.01	-0.02
Male	-0.01	-0.05	-0.01
IEP status	-0.05	0.05	-0.01
ELL status	0.03	0.02	0.01
Economically disadvantaged status	0.19***	-0.04	0.00
Race and ethnicity categories			
Black	0.03	-0.07	0.02
Asian	-0.23***	0.00	-0.04
Hispanic	0.18***	0.07	0.00
Other	-0.04	-0.03	0.04
Days between tests	-0.08	0.03	-0.03
Phonics pretest	0.06	0.11	0.05

Note. T-test statistical significance levels to test for differences between groups are indicated as ****p<.001, **p<.01,*p<.05.* Marginal statistical significance level indicated as †*p<.10.*

Table 2.3

Estimated Treatment (Standard Error) Impact: Children's Phonics Assessment with Selected Comparison Groups

	All possible comparison schools	Matched on student-level covariates	Refined balance on school-level covariates
Treatment	-0.01	-0.01	0.04
	(0.78)	(0.12)	(0.11)
Effect size	0.05	0.03	0.08
N	1,391	544	532

Note. Standard error in parentheses. Model 1 shows multilevel linear regression. Model 2 shows multilevel match with student-level covariates and multilevel linear regression. Model 3 shows multilevel match with student- and school-level covariates and multilevel linear regression.

Table 2.4

Fidelity of Implementation, Means (SD)

	All treatment classrooms ($N = 21$)	"High" fidelity classrooms ($N = 6$)	"Low" fidelity classrooms ($N = 16$)	Standardized difference
Survey				
Number of texts used	9.67 (3.86)	11.83 (2.32)	8.80 (4.07)	0.78
Used texts recently (1 = yes)	0.95 (0.22)	1.00 (0.00)	0.93 (0.26)	0.30
Proportion of students read text(s)	0.82 (0.26)	0.88 (0.21)	0.80 (0.29)	0.28
Perceived student engagement (out of 4)	3.55 (0.83)	3.83 (0.75)	3.43 (0.85)	0.49
Observations				
Total score	6.45 (3.01)	9.14 (1.36)	5.44 (2.84)	1.23**
Text matches phonics instruction	0.91 (0.29)	1.00 (0.00)	0.88 (0.34)	0.43
Texts available for students	0.53 (0.49)	0.67 (0.52)	0.48 (0.49)	0.33
Children reading texts	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.00
Lesson elements				
Lesson focuses on using phonics	0.25 (0.42)	0.47 (0.45)	0.15 (0.38)	0.76
Reviews and instructs in phonics	0.83 (0.32)	1.00 (0.00)	0.83 (0.39)	0.52
Connects to content	0.78 (0.43)	0.80 (0.45)	0.77 (0.44)	0.07
Introduces new words with word structure	0.4 (0.51)	0.83 (0.41)	0.11 (0.33)	1.42**
Introduces high frequency words with word structure	0.31 (0.48)	0.50 (0.58)	0.22 (0.44)	0.58
Prompts readers with reminders of phonics knowledge	0.62 (0.43)	0.94 (0.14)	0.48 (0.44)	1.08*
Prompts readers with strategies (e.g., say each sound)	0.63 (0.43)	0.88 (0.17)	0.52 (0.46)	0.85†
Students use strategies/skills	0.57 (0.47)	0.83 (0.41)	0.52 (0.46)	0.81†
Most of lesson is reading	0.50 (0.52)	0.75 (0.50)	0.40 (0.52)	0.67

Note. Sample included all treatment teachers. T-test statistical significance levels to test for differences between groups are indicated as $^{**}p<.01, ^{*}p<.05$. Marginal statistical significance level indicated as $^{\dagger}p<.10$.

Table 2.5

Association between Implementation Fidelity and Treatment, Children's Gains in Phonics

	Model 1	Model 2	Model 3
High quality	0.01	0.02	0.23†
	(0.13)	(0.03)	(0.13)
Effect size	0.08	0.00	0.05
N	381	332	332

Note. Standard error in parentheses. Model 1 shows association between higher fidelity and outcome. Model 2 shows association between score on observational fidelity scale and outcome. Model 3 shows association between using in-the-moment phonics instruction and outcome. Covariates included child-, class-, and school-level variables as described in-text. Marginal statistical significance level indicated as $^{\dagger}p<.10$.

Figure 1

Orthographic mapping

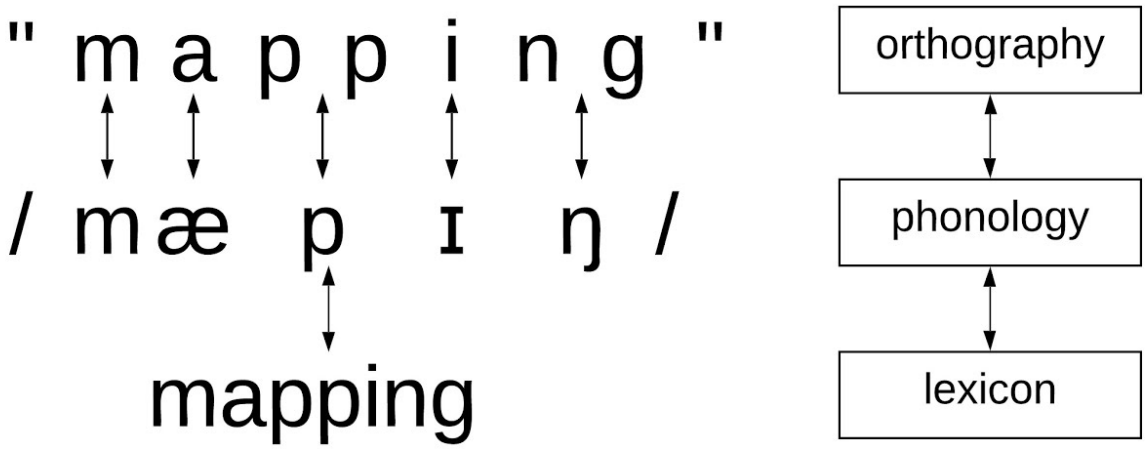


Figure 2

Beyond Decodables Text Criteria

Note. Graphic representation of some aspects of Beyond Decodables, supporting word recognition, word understanding, and text comprehension.

Chapter 3 A Preliminary Evaluation of Freedom Schools Literacy Academy: A Culturally Responsive Summer Literacy Program

Abstract

This study offers a preliminary evaluation of the Freedom Schools Literacy Academy, a summer program designed to support Black elementary schoolers' reading and racial identity development. Eighty-three kindergarten through second-grade children who participated in the program took assessments of listening comprehension, word recognition, oral reading fluency, and racial attitudes before and after the four-week virtual program. Children participated in phonics lessons with multiple criteria texts, culturally responsive read-alouds, and motivational activities in small groups via video conferencing (due to the COVID-19 pandemic). Results indicated statistically significant gains in listening comprehension, word recognition, oral reading fluency, and racial attitudes, across child characteristics and initial word reading abilities. Findings give positive support for Freedom Schools Literacy Academy and offer suggestive evidence in favor of early literacy programs that include research-based reading instruction along with attention to children's racial identity development.

U.S. public education largely fails to meet the needs of Black children. On the 2019 NAEP Reading assessment, only 18% of fourth-grade Black children scored proficient or above, with no significant improvement since 2011 (NAEP, 2019). Notably, the majority of teachers (93%) are non-Black (National Center for Education Statistics, 2020). Research has shown both that non-Black teachers often hold low expectations for the academic achievements of Black students (Dusek & Joseph, 1983; Gershenson et al., 2016; Tenenbaum & Ruck, 2007) and that low teacher expectations are detrimentally related to student achievement (Gershenson & Papageorge, 2018; Wang et al., 2018). Furthermore, a myriad of systemic challenges amounts to routine denial of access to resources, experiences, and more rigorous coursework for Black children (Chambers, 2009). Some of these systemic issues, which some researchers postulate are due in part to a cultural mismatch between Black students and schools, include overrepresentation in disciplinary practices, underrepresentation in rigorous academic tracks, and overrepresentation at under-resourced schools (Chambers, 2009; Kramarczuk Voulgarides et al., 2017). In recent times, during the COVID-19 pandemic, anecdotal reports indicate that some parents of Black children are reluctant to send their children back to in-person learning due to the respite from structural racism and racialized micro-aggressions experienced during distance/virtual learning (Anderson, 2020).

Evidence from research and practice demonstrate the possibility for another way for schooling to proceed for Black children. Scholars such as Ladson-Billings (1992), Gay (2010), Howard (2001), and Paris (2012) have introduced and advocated for teaching and pedagogies for Black children that prioritize honoring children's cultural backgrounds, experiences, and knowledge. Ladson-Billings (1992) described culturally relevant teaching (see Literature Review

for a more in-depth definition) as the necessary context for Black students to maximize learning, using cultural knowledge as a basis for learning and strength. Research and theory indicate that culturally relevant teaching is likely to support students' academic achievement and identity development (Morrison et al., 2008). In practice, groups such as the Children's Defense Fund have created culturally relevant programing to support children's academic and socioracial needs. The Children's Defense Fund Freedom Schools®, created in 1995, aims to provide a reformed and improved educational experience that emphasizes positive cultural messages for those children who have been traditionally marginalized in public schools (Jackson, 2011; Jackson & Boutte, 2009). A pretest/posttest study found that children participating in CDF's Freedom Schools® model did experience gains on an informal reading inventory over time (Lara-Cinisomo et al., 2020), suggesting that such programming may be both practical and supportive of reading development.

In the legacy of other Freedom Schools movements and programs, and in continued response to the demonstrated need to create a new kind of educational experience in for children who are Black, the Center for Black Educator Development (CBED), with support from researchers at the University of Michigan, created the Freedom Schools Literacy Academy (FSLA). In contrast to typical school experiences, FSLA is a Black-centric summer literacy program for children who will be entering first through third grade. The program is designed to support high academic expectations and achievement in early reading within a culturally relevant context. The program's goals are to positively impact the early literacy achievement of Black children, give Black children access to high levels of literacy, and support children's racial identity development. The CBED further aims to motivate and train future Black teachers in

culturally responsive practices and literacy development through their participation as Servant Leader Apprentices (SLAs) in FSLA (see Intervention).

In summer 2020, due to the impact on schooling of the COVID-19 pandemic, the Center for Black Educator Development's plans for FSLA's curriculum, format, and evaluation changed dramatically in a matter of weeks. While maintaining the program's primary goals, the shift to a virtual/distance format led to rapid development of curricula, selection of published curricula that could support the program's goals, and a light-touch evaluation design. In this paper, I describe a preliminary investigation into this novel program, examining children's gains in literacy and racial identity development, during the COVID-19 pandemic in a virtual summer literacy academy.

The findings from this study contribute to multiple literatures. First, this study is among a small number of studies attempting to quantify the impact of culturally relevant practices on elementary literacy outcomes. Second, in acknowledging the dual importance of academic and socioracial outcomes, this study contributes a unique, though cursory, look at young children's budding understandings of their race. Third, this study adds to the research on summer reading interventions aimed at supporting academic performance for children, especially some of whom we may anticipate could experience summer learning disparities (Alexander et al., 2007; Dumont & Ready, 2020). Finally, the program and this study add to the distinctively timely, but limited, research on the affordances and constraints of virtual/distance literacy learning in the elementary years.

Review of Literature

Culturally Relevant, Responsive, and Sustaining Early Literacy Education

Attempts to improve achievement on standardized assessments have often led to the creation and use of standardized curricula, especially in schools serving marginalized populations (Teale et al., 2007). Market research found that 78% of K-2 teachers surveyed in a nationally representative sample use a core reading program from one of a few major curriculum publishers (Meaney et al., 2017). Typically, this type of curriculum does not support teachers in adapting or differentiating for culturally responsive or sustaining teaching (Cummins, 2007).

In contrast, a priority of culturally relevant, responsive, and sustaining teaching is recognizing and honoring children's cultural backgrounds and experiences as assets for learning, which is difficult to do with materials that are designed to be the same for all children and communities. The terms culturally relevant, culturally responsive, and culturally sustaining, followed by pedagogy, teaching, or education, are all intricately linked and tend to be used interchangeably, despite some key distinctions. I discuss these distinctions in the following paragraphs.

Culturally relevant pedagogy is teaching with a social justice framework that privileges students' experiences and cultures to allow students from minoritized backgrounds to experience academic success, cultural competence, and critical consciousness (Ladson-Billings, 1995). Culturally relevant pedagogy recognizes and affirms students' cultural backgrounds and experiences as assets for learning that enhance students' ability to succeed (Morrison et al., 2008).

Culturally responsive teaching, originally described by Geneva Gay, "uses the cultural knowledge, prior experiences, frames of reference, and performance styles of diverse students to make learning encounters more relevant to and effective for [children]" (Gay, 2010, p. 31). Culturally responsive teaching emphasizes a social justice orientation and strengths-based

framework of children. According to Gay (2010), culturally responsive teachers strive to (a) socially and academically empower students; (b) engage students' cultural knowledge, experiences, contributions, and perspectives in a multidimensional manner; (c) validate every student's culture; (d) comprehensively educate each child socially, emotionally, and physically; (e) transform instruction, assessment, and curriculum; and (f) emancipate students from oppressive ideologies.

Culturally sustaining pedagogy (Paris, 2012) furthers both culturally relevant and culturally responsive pedagogies by insisting that teaching offer access to dominant cultural competencies while sustaining the linguistic and cultural practices of children and their communities. Culturally relevant, responsive, and sustaining teaching, therefore, aim to support children in accessing dominant cultural, linguistic, and academic success, while sustaining, honoring, and engaging children's pluralistic knowledges, culture, and power and supporting children in understanding the world through a critically conscious frame (Dahir, 2019).

Along with a larger body of descriptive and observational studies, a small body of quantitative research, mostly in secondary school, suggests that culturally responsive and other pedagogies are likely to support students' academic achievement and identity development (e.g., as synthesized by Aronson & Laughter, 2016 and Morrison et al., 2008). As of yet, however, there is relatively little empirical evidence that culturally responsive, relevant, and/or sustaining practices lead to higher academic outcomes in early elementary literacy instruction.

Although there is relatively limited research specifically investigating the impact of culturally responsive, relevant, and sustaining practices as a whole on elementary literacy, there is some research that supports the culturally responsive tenet of engaging students' cultural knowledge, experiences, contributions, and perspectives to improve literacy outcomes. In a small

quasi-experimental intervention study, Bui and Fagan (2013) found that using texts that reflected children's cultural experiences along with a series of research-based techniques to teach reading to fifth graders (N = 49) did lead to gains in children's reading, but not beyond reading other texts. Children's scores on the reading comprehension measure, however, did indicate a trend in favor of culturally responsive materials. In an experimental study, Bell and Clark (1998) found that elementary Black children (N = 109) had improved listening comprehension (F = 8.59, p < .01) about a text that focused on Black characters and African American cultural themes compared to texts with White characters and Euro-American themes. Together, these two studies point towards the possibility that when Black children's cultural knowledge is reflected in texts, they may comprehend texts better, which could lead to improved reading outcomes over time.

Several studies aim to describe how culturally relevant, responsive, and sustaining pedagogies and texts may encourage young elementary schoolers' identity development and motivation to read. Cartledge and colleagues (2016) described Black first and second graders' (N = 50) ratings of 30 texts, developed based on interviews and observations to specifically relate to the background and interests of the readers and to affirm the readers' racial identities. Children overwhelmingly rated the texts positively (93% positive ratings), giving reasons related to identity, enjoyment, and learning. Children's stated interest in reading culturally responsive texts may provide evidence of the potential for these types of texts to motivate children to read. Similarly, in another case study of three elementary schoolers, Piper (2019) observed that Black children interacting with civil-rights oriented read alouds in a Freedom Schools program demonstrated high motivation to read. These children also made comments that indicated positive understandings of their racial identity in relation to read alouds (Piper, 2019). Souto-

Manning (2009) also found that even first graders could be supported in discussions that fostered a budding critical consciousness within culturally responsive literacy lessons.

This small research base supports the possibility of positive reading gains when teachers enact culturally relevant, responsive, and/or sustaining practices. In particular, this research points to two culturally relevant, responsive, and/or sustaining practices that may impact literacy achievement. First, well-planned and thoughtful culturally relevant, responsive, and/or sustaining pedagogical stances and teaching practices, using research-based materials, may support literacy learning (e.g., Bui & Fagan, 2013). Second, culturally relevant texts may support children's engagement and comprehension and their racial identity development (e.g., Bell & Clark, 1998; Piper, 2019; Souto-Manning, 2009).

Summer Learning

Research over the last several decades continually points to summer as one source of opportunity and thus academic achievement disparities, and these disparities may disproportionately impact for children of color and children from lower-socioeconomic status communities (e.g., Alexander et al., 2007; Cooper et al., 1996; Entwisle et al., 1997; Hayes & Greather, 1983). Recent research points to far more complex relationships between summer, children's demographics, and learning outcomes (Dumont & Ready, 2020; Kuhfeld, 2019; Quinn, 2015; von Hippel et al., 2018). Studies by Dumont and Ready (2020), Quinn (2015), and von Hippel and colleagues (2018) all indicate that a researcher's chosen statistical model strategy and data source(s) may lead them to concluding that differences in achievement between groups may be mostly attributable to inequities in preschool, summer, or in-school, therefore calling into question exactly when and how these differences occur. These recent findings, however, still indicate that there are differences in academic achievement (in particular, reading) due to

differences in summer experiences (Kuhfeld, 2019) and confirm that the summer months may widen or at least not lessen inequalities in opportunities for children from lower socioeconomic status communities (Dumont & Ready, 2020).

Learning disparities related to summer may be especially consequential in reading in the early grades. In these earliest years of school, children need to become proficient readers, a challenging task if children are not reading during three months of the year. In particular, differences in reading achievement in these early years related to the summer may be exacerbated among children from lower socioeconomic backgrounds, children who are Black, and children who are male (Alexander et al., 2007; Quinn & Le, 2018; Slates et al., 2012).

Alexander and colleagues (2007) argue that summer learning differences by socioeconomic status in the early years account for a large proportion of differences in achievement in high school. Differences in reading at the beginning and end of the summer months between higher and lower socioeconomic status children are often attributed to changes in children's environment and access to resources. During the summer, children, especially from lower socioeconomic communities, may not have equal access to appropriate books or summer programs. Although not to a degree that would equalize opportunities, researchers, policy makers, and educators have created a multitude of summer programs to address the needs of low-income children in summer reading. Home and classroom-based programs are both effective ways to support children's reading growth in the summer months. Importantly, in a meta-analysis of summer reading programs, Kim and Quinn (2013) found that research-based summer reading programs in classrooms positively impacted children's reading comprehension and fluency and decoding above the positive impact of other classroom-based summer programs. In particular, Kim and Quinn (2013) found that summer programs targeting children exclusively

from lower socioeconomic households supported greater gains in reading than programs with a more socioeconomically diverse group of children.

Using data from the Early Childhood Longitudinal Study, Quinn and Le (2018) found that achievement differences between Black and White children were driven by increasing differences in the summer after kindergarten and into 1st and 2nd grade. Differences in Black and White children's reading, as typically measured by standardized tests of achievement, may be attributed to a myriad of factors, including systemic racial discrimination, comorbid socioeconomic challenges, and peer/family factors (Hung et al., 2019). Researchers, activists, and educators have created summer programs designed to address the needs of Black children. As previously described, a preliminary evaluation of one such program (CDF Freedom Schools® model) found that students (N = 784) who participated showed improvement (d = 0.4) on a reading inventory over the six-week program (Lara-Cinisomo et al., 2020). Children in kindergarten through second grade made the least gains, which may be due to the programmatic decision to not provide instruction in phonics or word reading. This study's outcomes suggest that a Freedom Schools model that combines aspects of CDF Freedom Schools® model and research-based instruction in phonics and word reading may support children's early literacy development.

Research-Based Early Reading Instruction

Research suggests that summer reading programs can positively impact a range of reading outcomes, but research on these programs tends to focus on the impact on phonics, fluency, and comprehension (Kim & Quinn, 2013). More broadly, reading interventions often do boost children's reading skills, both in the short and the long-term (particularly for older students in comprehension; Suggate, 2016). Short supplemental reading interventions (under 20 hours of

instruction), particularly in phonics, can also led to meaningful impacts (Allor & McCathren, 2004; Hatcher et al., 2004; Pericola Case et al., 2010), with interventions as short as eight hours leading to gains for children (Berninger et al., 2000). Further, specifically relevant to this sample, in a large (N = 552) experimental study of a voluntary at-home summer reading program using research-based teaching techniques, Kim (2006) found the greatest positive impacts for Black students. Taken together, these areas of research support the theoretical feasibility of the impact of a short, supplemental, and summer reading program that emphasizes research-based teaching techniques. In the following two sections, I overview some research behind the two major literacy-specific components of the program, phonics instruction and comprehension strategy instruction.

Phonics and Decoding Instruction

There is widespread recognition that systematic and explicit phonics instruction is an efficient and effective way to help teach word reading and is either beneficial or critical for the majority of children to learn to read (e.g., Henbest & Apel, 2017; National Reading Panel, 2000; Torgerson, et al., 2018). Phonics instruction that emphasizes the application of phonics knowledge to reading and writing tasks has the potential be the most useful, as the goal of phonics instruction is to help children read and write.

One way to teach phonics within an applied reading context is through texts that are have a high proportion of decodable words (often called *decodable texts*). Decodable texts are texts that contain a high proportion of words that children can read based on the grapheme-phoneme relationships and high frequency words they have learned. Theoretically, decodable texts complement systematic, explicit instruction in phonics by giving children the chance to practice using grapheme-phoneme correspondences in reading actual texts to contextualize and generalize

these skills, which researchers suggest is critical to reading success (Rupley, Blair, & Nichols, 2009; Stein, Johnson, & Gutlohn, 1999; Taylor, Pearson, Clark, & Walpole, 2000).

Empirical research on the impact of decodable texts alone is limited; however, more broadly, research points to positive impacts of decodable text reading within a connected phonics intervention for a wide range of readers (Allor et al., 2020; Beverly et al., 2009; Cheatham et al., 2014; Chu & Chen, 2014; Fien et al., 2015; Hiebert & Fisher, 2016; Jenkins et al., 2004; Juel & Roper-Schneider, 1985; Mesmer, 2005; Pericola Case et al., 2010). Contemporary research extends beyond simply investigating the impact of the decodability of texts on readers' outcomes and instead reflects research on a multitude of text factors that affect reading development to attempt to create the best possible texts for early readers (called multiple-criteria texts). Research suggests that reading multiple-criteria texts, with attention to decodability, other word-level factors (e.g., repetition, high frequency words), and meaningfulness, may support developing readers (Allor et al., 2020; Cheatham et al., 2014). One recipe for effective word reading instruction, therefore, may be explicit and systematic phonics instruction with many opportunities for application in multiple-criteria texts.

Interaction Read-Alouds

Word-reading instruction is not the only way to improve reading achievement. A recent study found that low socioeconomic status children may not experience gains or stability in reading comprehension when involved in a summer program that focuses on decoding only (Nicholson & Tiru, 2019); thus, this component is essential to include. One way to facilitate comprehension development is through interactive read-alouds. Interactive read-alouds include research-based strategies for engaging children actively in before, during, and after reading to co-construct meaning. Key in interactive read-alouds is making high-quality book choices and

supporting children's language and understanding (Baker et al., 2013; Lennox, 2013). As discussed above, one key component in engaging and supporting the comprehension of Black children may be reading aloud culturally relevant texts.

Another way to support children's comprehension development in interactive read-alouds is through comprehension strategy instruction. Comprehension strategy instruction has also been shown to help students develop skills and strategies that support comprehension development (Duke et al., 2011). Comprehension strategy instruction includes explicit instruction around the specific mental actions that may help a reader better understand a text (Shanahan et al., 2010). An enormous body of research has found that teaching children to use comprehension strategies positively impacts comprehension (e.g., Morrow et al., 1995; Spörer et al., 2009). FSLA instruction focuses on supporting children through explicit strategy instruction embedded in interactive read alouds. Children learned to apply two comprehension strategies supported by research: activating background knowledge and retelling.

Activating background knowledge (encompassing predicting), may help children understand a text by supporting their ability to make inferences and connections to prior knowledge and is often considered an essential part of supporting children in reading comprehension and knowledge development (Brown et al., 1995; Cervetti & Hiebert, 2015; McClure & Fullerton, 2017). Further, in order to have relevant background knowledge for a child to activate, a text should ideally be matched to a child's background (culturally relevant) and interests. A recent study found that children with higher levels of background knowledge relevant to the text made more appropriate inferences than children who learned new knowledge to access a text (Kaefer, 2020), suggesting that the combination of culturally relevant texts and activating prior knowledge may support comprehension of a particular text.

Retelling helps a child organize and describe a text. Several studies find that using a retelling strategy aids in comprehension across the early elementary grades (e.g., Hagaman et al, 2016; Morrow et al., 1990). Morrow and colleagues found that kindergarten children engaged in literature experiences including story retelling, with references to story elements, had greater growth on standardized measures of reading achievement than a control group. More recently, Hagaman and colleagues (2012) demonstrated that a strategy to retell or paraphrase small sections of a text supported third graders' reading comprehension. Together, explicit instruction and scaffolded support in these two comprehension strategies (a pre-reading routine to activate background knowledge and a retelling routine) may be one way to support children's comprehension development.

Reviewed Literature in Relation to FSLA

Taken together, the reviewed literature highlights the clear need filled by the Freedom Schools Literacy Academy (FSLA) and the research basis for the components of the program. First, FSLA is a summer program due to the theoretical, empirical, and practical likelihood that summer is likely one source of the disparities in learning outcomes for children of color and children from lower-socioeconomic backgrounds (e.g., Alexander et al., 2007; Cooper et al., 1996; Entwisle et al., 1997; Hayes & Greather, 1983). Improving summer learning may be especially consequential for children from low-socioeconomic status communities, boys, and younger children. Further, summer programs can have differential impacts on particular groups of children, suggesting additional program creation and evaluation is needed to develop programs that support a wider range of children (Kim & Quinn, 2013). Second, in acknowledging that Black children's needs may not be met by business-as-usual school, FSLA employs culturally relevant, responsive, and sustaining teaching pedagogies, which recognize

children's cultural backgrounds as assets for learning and springboards for academic success (Dahir, 2019; Gay, 2010; Ladson-Billings, 1995; Morrison et al., 2008; Paris 2012). These pedagogies, paired with research-based literacy instruction, may support children's literacy learning and racial identity development (Bell & Clark, 1998; Bui & Fagan, 2013; Cartledge et al., 2016; Piper, 2019; Souto-Manning, 2009). Third, FSLA includes instruction likely to impact children's word reading and comprehension; namely, a phonics with multiple-criteria text program and interactive read-alouds.

The Current Study

To add to the literature on culturally relevant, responsive, and sustaining early literacy instruction, I addressed the following research questions in my investigation of the Freedom Schools Literacy Academy:

1. Did children who participated in the Freedom Schools Literacy Academy show gains on measures of listening comprehension, word reading, and positive racial identity?
2. Are gains in word reading, listening comprehension, and racial identity associated with children's characteristics (age, gender, and socioeconomic status)?

Methods

Participants and Setting

The study sample consisted of 83 children who participated in Freedom Schools Literacy Academy (FSLA) in Summer 2020 and whose parents consented for their data to be used in research (consent rate = 79%; see Table 1 for demographic information). Children were residents of eight different states across the United States, with a majority (75.9%) of children coming from Pennsylvania. On average, participants were 6.78 years old (*SD* = 0.98), with approximately one-third of children in each grade band (rising first grade [i.e., entering first

grade in the fall] $N = 30$; rising second grade $N = 26$; rising third grade $N = 27$). Most participants (94%) identified as Black/African American. Participants who did not primarily identify as Black/African American identified as primarily Latino/Hispanic (2.41%) and Other Races (3.61%). About 65% of participants attended public schools. Of those who attended public schools, over half (56.9%) attended high-poverty schools (defined as "public schools where more than 75% of the students are eligible for free or reduced-price lunch;" NCES, 2020).

Intervention

The Freedom Schools Literacy Academy (FSLA) was created by the Center for Black Educator Development, based in Philadelphia, PA, and researchers at the University of Michigan in the spring of 2019. FSLA launched as a pilot of a new model of Freedom Schools in the summer of 2019. The program was created to expand on pre-existing Freedom Schools models, such as the Children's Defense Fund Freedom Schools® and the Philadelphia Freedom School, and to continue a commitment to culturally responsive literacy education, train new Black educators, and support the literacy learning of participants. In the summer of 2020, due to the impacts of the COVID-19 pandemic, the program shifted to a four-week (30 hours of programming, including 20 hours of direct literacy instruction) virtual program designed to meet the same goals. The program took place over Zoom, Inc.'s videoconferencing software. The Center for Black Educator Development gave children and staff without access to technology in the Philadelphia metropolitan area a tablet or computer for the program. All participants were required to have internet access.

Servant Leader Apprentices

The instructors in the program were college-aged students and recent graduates interested in pursuing careers in education, called *Servant Leader Apprentices* (SLAs). In addition to

supporting children's literacy development, a goal of the Center for Black Educator Development is motivating and training potential teachers in culturally responsive practices and literacy development. SLAs were recruited and interviewed by program staff. All SLAs identified as Black. Prior to the program's start, the 26 college-aged SLAs participated in 24 hours of training and coaching in the curriculum, classroom management, literacy development, and Black-centered education from four program literacy coaches employed by the Center and two research partners, from the University of Michigan (the author) and Southern Methodist University (a developer of *Friends on the Block,* described in the following paragraph). Throughout the program, SLAs also attended approximately six hours of additional professional development and seminars focused on deeper understandings of the curriculum and assessments, experiences of Black educators, and culturally responsive education. Each Servant Leader Intern also received personalized coaching on implementation of the summer program components and integration of culturally responsive practices from a coach who observed and supported their teaching. Servant Leaders were observed for at least one lesson (30 minutes) per week, met with a coach for up to an hour each week individually, and met in groups with coaches for approximately 30 minutes a day. All coaching, training, assessment, and instruction occurred in virtual settings.

Instructional Components

FSLA had three major components: a) an explicit, systematic phonics program with accompanying multiple-criteria decodable texts called *Friends on the Block*; b) a culturally relevant interactive read aloud curriculum with comprehension strategy instruction; and c) a culturally sustaining, Black-centric morning meeting called *Harambee*. Each component is discussed below. All children participated in all components.

Friends on the Block. The explicit, systematic phonics program (seven 30-minute lessons per week; 14 hours total) with accompanying multiple-criteria decodable texts used during FSLA was a revised version of *Friends on the Block* (Allor et al., 2019), an early literacy curriculum initially designed for children at the beginning stages of reading. Although initially designed for children with intellectual disabilities, the program developers envisioned that the approach used in the intervention could be used to support any beginning reader. At its core, Friends on the Block is a highly systematic, explicit phonics curriculum with embedded opportunities to practice reading in multiple-criteria texts. As explained earlier, systematic, phonics instruction is an essential component of reading instruction for beginning reading instruction (de Graaff et al., 2009; Henbest & Apel, 2017; Torgerson et al., 2018). To date, there have been two published case studies, using a pretest/posttest design with a total of 18 children, of Friends on the Block that demonstrate that students show gains in reading following use of the program and that the program is feasible for a range of teachers to implement with fidelity (Allor et al. 2018; Allor et al., 2020). Friends on the Block was selected due to its user-friendly lessons, materials easily adaptable for a virtual setting, and theoretical support of some key features of the curriculum that support beginning readers.

Prior to the program's start, trained assessors gave children a word recognition assessment connected to the content of Friends on the Block to place children into small groups (3-5 children) based on their needs. Children's groups changed weekly in order to maximize the opportunity for differentiated instruction in the pace and content of lessons. This decision was made at the suggestion of the program's developers. In each 30-minute lesson, children participated in explicit instruction and practice in phonemic awareness, phonics, and high-frequency words. The explicit phonics instruction begins with basic alphabetic knowledge and

moves to more complex within-word patterns (e.g., r-controlled vowels and other vowel patterns). In each session, after practicing these skills in isolation, children read a text including words with these phonics patterns. The Friends on the Block texts include a high percentage of words decodable based on what students would have been taught to that point in the curriculum, settings likely to be familiar to young children, high frequency words that have been taught to that point, natural syntax, repetition of individual words, words likely to be familiar to young children, and cumulative practice. In order to increase the meaningfulness of the texts, each text has a forward and many have helper text read by the teacher that add context to the decodable reading. After reading, children participated in a series of games to reinforce their phonics skills. For example, children played a bingo-like game in which they identified a written word after hearing it aloud or seeing a picture of an object.

Some children ($N = 32$) tested at or above the ceiling of the Friends on the Block assessment at pre-test. We adjusted lessons for those children to include more instruction in more complex phonics skills and texts as well as more fluency practice through repeated readings of multiple texts across the week (inspired by Kuhn et al., 2006).

We adjusted the Friends on the Block curriculum to work in a short, virtual program and to be used in a culturally responsive manner. First, we transferred games into an online format using Google Drawings. Next, we transferred all lesson materials into PowerPoint slides. The texts are e-books, so these minor adjustments allowed the program to be virtual-learning friendly. Second, we used assessment-based, flexible groupings and regrouped each week to accelerate progress through consistent assessment and response to children' needs and strengths. Third, teachers were trained to engage in culturally responsive practices while using this program, including explicitly skipping or calling out texts that did not feel relevant to children

and honoring children's language while reading and speaking (e.g., by respecting children's dialects as appropriate differences, not as errors).

Interactive Read Alouds. The second component of the program was twelve 30-minute (6 hours total) interactive read aloud sessions. These read-alouds featured texts with Black authors, characters, and/or narrators available through an online platform called Storyline Online, which streams videos of actors reading children's books along with the text, illustrations, and occasional animations. Servant Leader Apprentices (SLAs) first engaged children in activating background knowledge. Then, while "reading," they paused read-aloud videos at key points to ask questions and support children's vocabulary learning. Each lesson also included explicit instruction and practice in retelling. During each lesson, children learned new vocabulary, made predictions, answered comprehension questions, retold the story, and engaged in a connected activity, such as drawing and writing about their similarities to a character. An experienced elementary literacy coach and I wrote the interactive read aloud lessons to align with research and the program's goals for positive racial identity development.

All the texts featured Black authors, characters, and/or narrators. This choice aimed to support children' racial identity development and comprehension. As explained earlier, research suggests that children may have improved comprehension of texts that reflect their cultural knowledge, themes, and racial backgrounds (Bell & Clark, 1998; Bui & Fagan, 2013; Cartledge et al., 2016; Piper, 2019; Souto-Manning, 2009).

As previously noted, FSLA explicitly aimed to allow children to see themselves represented in books (i.e., looking into a mirror; Bishop, 1990) as many other educational experiences for the Black children participants were unlikely to provide this experience. In

addition to supporting comprehension, we hoped read-alouds would support children's positive racial identity.

Within each read aloud, SLAs engaged children in explicit instruction and practice through a gradual release model (Pearson & Gallagher, 1983) in discussing story elements, activating prior knowledge, making connections, and retelling. The lessons also featured questioning to support children's deeper understandings of the texts. The questions centered around four themes, developed to help support children' positive understandings of themselves and their community: (a) I am unique and special; b) My family and community are unique and special; c) I can achieve any dream; d) My voice deserves to be heard.

Harambee. Each day of FSLA began with a 30-minute (10 hours total) motivational experience called *Harambee* (Kiswahili; translated as *let's pull together*). Harambee is a component of other Freedom Schools models and uses cheers, chants, community recognitions, and brief culturally relevant read alouds to engage and excite children. In Harambee, read-alouds are not interactive, but instead focus on introducing children to a community member who reads the story aloud to encourage children to recognize that reading is important, no matter who they are or what profession they would like to have as an adult.

Procedures

The Institutional Review Board of the University of Michigan deemed this study exempt from oversight through the exemption for educational research involving normal educational practices**.**

Recruitment

Program staff invited all enrolled children to participate in this study. FSLA staff contacted parents and explained the research project, participation, and consent through email

during the assessment period. Participation in the research study did not impact participation in the program itself. Eighty-four children (out of 105) and their parents consented to participate. One child did not attend the program. The analytic sample is 83 children.

Assessment

All assessments were part of internal evaluations for FSLA. Thus, assessment procedures were based on the program's needs and constraints. Fourteen trained assessors administered the pre-test tasks (word recognition task, listening comprehension task, and racial and social identity scale) during the two weeks before the program's start. SLAs (also trained in administering the assessments) assessed children who hit the ceiling of the pre-test word recognition task on an oral reading fluency task during the first day of the program ($N = 32$). On the final day of the program, SLAs administered the post-test tasks (word recognition task or oral reading fluency task, listening comprehension task, and racial and social identity scale). SLAs assessed the students in their group at that point in the program. Two trained assessors (from the original 14) administered six make-up post-tests (including all tasks above) within four days of the program ending.

Measures

Listening Comprehension

The listening comprehension task was a shortened version of the Narrative Language Measure (NLM) Listening (a subtest of the CUBED assessment; Petersen & Spencer, 2016). The interrater reliability coefficient for NLM Listening is between 0.82-0.96 and the predictive validity coefficient with Northwest Evaluation Associations' Measures of Academic Progress is 0.43 (Petersen & Spencer, 2016). Recent work demonstrates the content validity of the NLM Listening comprehension task and reading comprehension (Petersen et al., 2020). The adapted,

shortened version focused on retelling narrative story elements. After listening to a short, grade-level narrative story, children retold the story. Children's retellings were scored based on inclusion and description of characters, settings, problems, solutions, feeling or descriptive words, and a logical sequence. Listening comprehension scores are reported as a percent out of a total possible score of 12. All children took the listening comprehension assessment.

Word Reading

Depending on skill level, children either took the word recognition assessment ($N = 46$) or the oral reading fluency assessment ($N = 32$).

Word Recognition. The word recognition assessment is part of the Friends on the Block curriculum (Allor et al., 2019). The assessment consists of 5-30 words per level across 12 levels (245 words total). Each level assesses words that include a combination of high-frequency words and words that are decodable based on a reasonable developmental trajectory in phonics. For example, the assessment begins with reading high frequency words and consonant-vowel-consonant words and moves to reading words with consonant digraphs, long vowel patterns, and r-controlled vowels. At each level, during the course of *Friends on the Block* instruction, children learn and practice reading all the words that appear in that level's task. Thus, the assessment is measure of children's learning of taught words. At this time, no psychometric information is available. However, examining children's accuracy in reading lists of individual words of increasing difficulty is an approach that has been used in the past (e.g., Torgesen et al., 1999, see "test of word reading efficiency"). Children's scores are reported as raw scores (levels, 0-12) and percentages (out of 12).

Oral Reading Fluency. Children who tested at or above ceiling ($N = 32$) on the word recognition task (see above) during the pre-test were then assessed using the Oral Reading

Fluency (ORF) subtest of Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminiski, 2002). DIBELS ORF is used across the country in many populations both in research and for progress monitoring in elementary reading classrooms. The median coefficient on alternate form, concurrent reliability for DIBELS ORF (words read correctly) in a face-to-face context is 0.93 (University of Oregon, 2020a) (no psychometric information about its use in a videoconferencing context is available, however, test developers recommend caution in interpreting remote test scores; University of Oregon, 2020b). The median coefficient for inter-rater reliability is above 0.99. DIBELS concurrent criterion validity coefficients for grades 1-3 range from 0.24-0.91 with the Iowa Assessment Total Reading and Word Analysis assessment (University of Oregon, 2020a). DIBELS is generally considered to be sensitive to intervention effects and has been used in exploring the impact of phonics interventions (e.g., a synthesis of research on programs for struggling readers found that about 8% of reviewed studies primarily used DIBELS as an outcome measure; Slavin et al., 2011).

Each child read for one minute. Children read passages appropriate for fall in their grade level in the 2020-2021 school year. SLAs received training on ORF scoring and practiced scoring adult readers and child recordings prior to assessment. SLAs wrote or recorded all errors and possible errors while listening to children's reading, and I scored each child's reading by calculating the words correct per minute (WCPM). ORF scores are reported as both raw scores (WCPM) and as nationally normed grade-level percentiles (University of Oregon, 2020a).

Racial Identity Scale

The Center for Black Educator Development's literacy coach team adapted the racial identity scale from Smith and colleagues' (2003) Racial Attitudes Survey (RAS). The RAS was originally created for upper elementary school Black students and was validated in this group

(Smith et al., 2003). The RAS measures three underlying factors of racial-ethnic attitudes: racial-ethnic pride, perception of racial barriers, and racial trust/mistrust (Smith et al., 2003). The coach team adapted the assessment to fit the age and racial backgrounds of children (see Appendix 2.A for assessment). In this adapted version, children were only asked questions related to racial pride. All children were first asked to explain the word *race* and identify their race. Children who understood the term *race* and their own racial identity were asked two additional questions about their attitudes and pride in their racial identity. The racial pride and identity scale is reported as scores on Likert scales (0-3 scale, where 0 represents "item not given based on initial racial identity question").

Child Characteristics from Administrative Data

Children's demographic information was collected as part of the program's entrance form. I created an indicator for whether a child identifies as Male (coded 1). As I was unable to ascertain family income directly through the intake form, I created two different proxies for socioeconomic status based on information collected through this form. Some research suggests that local income levels, in particular concentrations of poverty in a school or neighborhood, may predict student achievement more strongly than individual family income (Sampson & Sharkey, 2008). Thus, I created an indicator for whether median household income in the child's home zip code fell below the national median (coded 1; US Census Bureau, 2021) and whether the child's school is considered high poverty (75% of more of students are free- and reduced-price lunch eligible; coded 1; all private schools coded as 0).

Analytic Approach

Missing Data

In general, there was a low level of missing data across children's characteristics and pretest scores (from 0.00% to 9.40%). As shown in Table 1, the only child characteristic with missing values was child age (3.61% missing). As age is highly related with grade level, I replaced missing age data with the mean age for the child's grade level. There were no systematic difference between children with and without missing data on the predictors. As the rate of missingness was relatively low, I used complete case analysis in the main set of results.

I found that outcome data were missing at relatively low rates (2.4% to 9.8%) for the listening comprehension measure, racial identity scale, and word recognition measure. Differences in children's characteristics based on missingness on these outcomes were not statistically significant. There was a high level of missingness on the oral reading fluency posttest (recall that 32 children total were in the group that took this assessment; 25.00% of children were missing from the posttest). Children with missing oral reading fluency data were more likely to be younger (both by age and grade level) than children without missing data. One assessor gave the incorrect grade-level passage on ORF to five children, which accounts for a high proportion of the missingness (62.50% of those with missing data on ORF). Five additional children did not have full data on ORF for a variety of reasons. In the main models I present, I use complete case analysis, but in Appendix 2.B, I also present robustness check models using multiple imputation for the ORF task (as recommended by Cox et al., 2014). In the first model for ORF, I impute 100 data sets using multivariate normal regression and impute only pre-test data. In the second model, I impute 100 data sets using multivariate normal regression and impute pre-test and outcome data (van Ginkel et al., 2020).

Confirmatory Factor Analysis

In order to investigate the factor structure for the racial identity scale, I conducted confirmatory factor analysis with the three indicators. Based on the work by Smith and colleagues (2003), I defined the one construct with 3 indicators as racial identity and pride. I conduced the CFA with only pre-test answers ($N = 82$). There was excellent fit; however as the sample size is small, this represents only preliminary evidence that the adjusted scale may be appropriate for this age group (root mean square error of approximation, RMSEA 90% confidence interval, CI [0.00, 0.00]; comparative fit index, CFI = 1.00; Tucker Lewis Index, TLI = 1.00; standardized root mean square residual, SRMR = 0.00; Hu & Bentler, 1999). Cronbach's α for the racial identity and pride construct was 0.93 (above the benchmark of 0.80; Clark & Watson, 1995).

Descriptive Analysis and Gains

I used Stata/IC version 16 for Mac (StataCorp, 2019) for descriptive analysis and modeling. To answer the first research question (*Did children who participated in the Freedom Schools Literacy Academy show gains on measures of listening comprehension, word reading, and racial identity?*), I calculated means, standard deviations, and ranges for each measure at pre- and post-test. I used paired t-tests to determine whether gains were statistically significant.

Modeling Associations with Gains

To answer the second research question (Are gains listening comprehension, word reading, and racial identity and pride associated with children's characteristics?), I used residualized gains models in which I separately regressed each outcome on the pretest score and a key child demographic predictor in a series of models. These predictors included child's age, gender, and proxies for socioeconomic status. I report robust standard errors in all models. Due

to programmatic decisions about assigning children to SLAs, children's groups and SLAs were not stable, and therefore, I did not cluster standard errors by SLA.

Results

Research Question 1: Changes in Reading and Attitudes

Children made statistically significant gains on measures of listening comprehension, word reading, and racial identity and pride over the four weeks of the intervention. In listening comprehension, on average, children (N = 80) improved narrative retell scores by 16% over the four-week program (pre-test M = 0.67 SD = 0.23, post-test M = 0.83, SD = 0.20, t = 5.05 $p <$ .001; see Table 2).

In word reading, children made statistically significant gains on both types of assessment. For children who took the word recognition assessment (N = 46), children made statistically significant gains, improving their word recognition measure scores by an average of approximately 24% over the program (pre-test M = 0.31, SD = 0.26, post-test M = 0.55, SD = 0.24, t = 14.50, $p < 0.001$; see Table 2). In other words, children improved an average of approximately 2.83 levels (out of 12 levels) over the course of the four week, 35 hour program (with 14 hours focused on word recognition instruction). For example, children who could read words with regular consonant-vowel-consonant patterns (e.g., *mad* and *ran*) at the beginning of the program could, on average, read words with some common long vowel patterns (e.g., *gave* and *make*) by the end of the program.

Recall that children who had reached the ceiling on the word recognition assessment at pre-test were administered the oral reading fluency assessment. These children (n = 22) made statistically significant gains in oral reading fluency (words correct per minute), reading on average approximately 21 more words per minute over the four-week program (pre-test M =

69.63, *SD* = 28.49, post-test *M* = 90.18, *SD* = 26.48, *t* = 5.39, $p < 0.001$; see Table 2). On average, children began the program with oral reading fluency scores at the 41st percentile based on national grade-level norms (University of Oregon, 2020) and ended the program at the 60th percentile (pre-test *M* = 0.41, *SD* = 0.25, post-test *M* = 0.60, *SD* = 0.25, *t* = 5.32, $p < 0.001$; see Table 2).

Children also made statistically significant gains on the racial identity and pride construct. Children (*N* = 80) scored an average of 0.40 points higher on the 0 to 9 Likert scale items for the racial identity and pride construct (pre-test *M* = 1.74 *SD* = 1.18, post-test *M* = 2.14, *SD* = 1.12, *t* = 2.31 $p < 0.05$; see Table 2).

I also investigate the changes in racial identity and pride specifically for children who learned about their race during the program. Twenty-two children who did not know their race at the beginning of the program did know the word *race* and could identify their own race at the end of the program. Of these 22 children, 95% said they liked their race and 100% said they should be proud of their race at the end of the program. These results indicate that children who became more racially aware during the four-week program also gained a positive view of their own race.

Research Question 2: Associations between Changes and Children's Characteristics

Children's demographic characteristics were not associated with gains on measures of listening comprehension, word reading, and racial identity and pride (see Table 4). Living in a zip code with a median income lower (mean = 37,011) than the national median (62,843; U.S. Census Bureau) was statistically significantly associated with lesser gains on the word recognition measure (standardized difference = 0.58). Living in a zip code with a median income lower than the national median was also statistically significantly associated with lesser gains on

the oral reading fluency measure (standardized difference = 0.42), when controlling for other variables, while attending a high poverty school was statistically significantly associated with higher gains on the same measure (standardized difference = 0.25).

These results should be interpreted with extreme caution for three reasons. First, these divergent results call into question the validity of the chosen proxies for socioeconomic status. Second, these models have very low explanatory power due to the overall sample size. Third, as most children lived in lower income zip codes (81.93%), the comparison groups are drastically different sizes. Further, in the listening comprehension and racial identity models, the *R*-squared is low; these models only explained between 1% and 14% of the variance, suggesting changes may be due to other factors or the program itself. In both word reading measures, recall that only a portion of children took each subset (word recognition, $N = 46$; oral reading fluency, $N = 22$), so power to analyze these subgroups is very limited.

Discussion

This study was a preliminary investigation into the reading and racial attitudes changes associated with a novel summer literacy program in a virtual format that aimed to be culturally responsive and sustaining. Results indicated that children who participated in the program did make statistically significant gains on measures of word reading, listening comprehension, and racial attitudes. This study is the first to evaluate this new program and is one of the first investigations of a summer, virtual/distance-learning reading program for young children.

My findings add to research on the impact of literacy instruction, particularly in the summer. These findings are consistent with research on summer reading programs. Kim and Quinn (2013) found that research-based summer reading programs tested in experimental and quasi-experimental studies tended to have a moderate to large effect on reading comprehension

(*ES* = 0.38) and fluency and decoding (*ES* = 0.63), which is comparable to the significant gains on similar constructs of listening comprehension and word reading in the present study.

There are three important distinctions in this study's results compared to those reviewed by Kim and Quinn (2013). First, I found statistically significant gains in word reading and listening comprehension after just 20 hours of direct instruction and practice. By contrast, Kim and Quinn's (2013) meta-analysis found a positive main effect of resource-intensive summer programs (defined as (a) fewer than 13 students per class, (b) 4 to 8 hours of instruction per day, and (c) 70 to 175 hours of total instruction), but a non-significant effect for less resource-intensive programs. In this study, children participated in only an hour of direct instructional time each day (20 hours of direct instruction, 30 hours of program time), far less total time and daily time many other summer programs. This is not entirely unheard of as research has found that short, supplemental reading programs can impact children's outcomes (Allor & McCathren, 2004; Berninger et al., 2000; Case et al., 2010; Hatcher et al., 2006).

Though not "resource-intensive" by Kim and Quinn's standards (2013), FSLA was instructionally intensive in other ways, which may account for the gains associated with program participation. First, instruction occurred in small groups of 3-5 children, giving teachers the ability to support children individually. Second, the program used continuous progress monitoring and regrouping to differentiate instruction. These instructionally intensive moves that allowed for differentiation of instruction may account for some of the results of this short program. These results are consistent with a recent meta-analysis on differentiated literacy instruction (Puzio et al., 2020), which found that current studies of differentiation reading instruction find that children's letter-word reading outcomes improve after differentiated instruction.

A second distinction in this study compared to other studies on summer reading is that there were significant gains for children in both word reading and comprehension. About half of the studies reviewed by Kim and Quinn (2013) included only a measure of fluency and decoding or reading comprehension, but not both, while the present study included both word reading and listening comprehension measures and found effects on both. Moreover, the program targeted both word reading and comprehension development. The Friends on the Block curriculum was geared toward word reading. As previously noted, the combination of explicit phonics instruction along with consistent practice in multiple criteria texts may have supported children not only in word recognition (as demonstrated by the word recognition assessment), but also in oral reading in context (as demonstrated by the oral reading fluency assessment). The read aloud curriculum, in addition to supporting socioracial development, integrated explicit instruction in comprehension strategies, which may have contributed to children's gains in listening comprehension. The Friends on the Block curriculum also included some questions to support children's comprehension of the multiple criteria texts, giving children another chance to think about text meaning.

The third critical distinction from most other summer literacy programs is the gains on one measure of racial attitudes. Though additional research is needed to understand how or if the program caused these gains, they may be partially due to the program's intentional culturally responsive, relevant, and/or sustaining early literacy instruction. Further, although critics may point out that without more extensive observations of teachers, it is impossible to definitively say that the program lived up to its aim of providing a culturally sustaining experience, at a minimum it is clear that the trainings, ongoing coaching, and lesson plans aimed to support SLAs in delivering culturally responsive and sustaining instruction. Additionally, other research has

shown that teachers can take on more culturally responsive practices through professional development and coaching (Hilaski, 2020), as was offered in this program. In informal exit interviews, multiple SLAs, coaches, children, and their parents commented on noticing, enjoying, and finding comfort in culturally responsive elements of the program, giving some informal support to the SLAs' implementation. Thus, the findings presented in this paper do suggest a the possibility of culturally responsive literacy instruction. For this population, it is clear that children could make reading gains within a culturally responsive and sustaining summer literacy program. It is equally clear that, for this population, children could demonstrate some improvement in their budding racial attitudes, even within a program that spent 70% of its time in explicit phonics and decoding instruction.

Although I cannot say whether these results are generalizable to other implementations of the program or definitively whether these results are due to the program, they point to an important possibility for literacy researchers, policymakers, and practitioners: culturally responsive and sustaining teaching that use techniques, curricula, and materials aligned with research may support reading gains *and* children's identity development. Although further research is required, the present study preliminarily suggests that it is unlikely that cultural responsiveness approaches and reading development are mutually exclusive, despite limited attempts to integrate these ideas in research.

Another key finding in this study is the stability of gains for children, despite socioeconomic, gender, and age differences. Evidence from Kim and Quinn's (2013) meta-analysis suggest that gains are highest when programs only include children from low socioeconomic communities; however, this program included children from a range of socioeconomic communities and schooling backgrounds and children still all made similar gains.

Finally, a key finding from studying the Freedom Schools Literacy Academy is that gains for children can occur in the context of virtual reading instruction. Although undoubtedly interest in virtual learning for elementary schoolers will decline as the COVID-19 pandemic ends, it is helpful to have examples of virtual reading programs that "work" for future scenarios and continued use of virtual programming, for those that prefer it and, potentially, to supplement a need (e.g., in rural areas with limited access to schools, for non-White families seeking a less racist environment, or for specialized tutoring). This preliminary investigation shows that the Friends on the Block curriculum, interactive read alouds, and additional practices adapted for a virtual format can help young readers, even at the very earliest stages of reading.

Limitations

This study has a number of limitations. First, the study used a simple pre-post design and, therefore, results cannot be interpreted causally. Any estimated gains of children in this program could be attributed to a myriad of confounding factors, including that children whose parents were motivated to sign them up for this program could live in home environments in which they would have made considerable gains in literacy and racial identity over the summer regardless of whether they participated in the Academy. This design, without a clear comparison group, could also led to biased results as statistical regression to the mean in pretest-posttest may inflate observed gains. All results should be considered suggestive evidence that the program may be associated with reading and attitudes gains for a very particular population of children. Additional research with comparison groups is needed to determine whether the program supports a wide range of children above and beyond other programs.

Second, the program was, by necessity, implemented incredibly quickly. One result of the quick implementation was that there was not time to check interrater reliability on measures that

have subjective elements (namely, the listening comprehension measure). Another result of the need for immediate implementation of this program is that a phonics program, Friends on the Block, was used, despite not matching with all of the goals or needs of FSLA. In an effort to select a research-aligned, theoretically sound beginning reading program that novice teachers could implement, program staff had to sacrifice an ideal vision of texts and materials specifically designed for Black children, which could have contributed further to children's racial identity development. An additional limitation was the small sample size used to confirm the factor structure of the racial attitudes scale. Additional research is necessary to determine whether this measure is validly measuring this construct in young children.

Conclusion

Despite the preliminary nature of the present study, two factors mean preliminary evidence in this space may still be helpful. First, in this year of turmoil, it has been necessary to quickly understand how children's learning may be supported in virtual environments. This study, although limited, may help illuminate whether a virtual small group reading intervention focused on culturally responsive materials and interactions helped ensure that this is not a "lost summer" for children's learning. Second, there is limited information about how culturally responsive practices support early elementary schoolers, so this study is able to add suggestive evidence that even short, imperfect implementations of culturally responsive, relevant, or sustaining practices may help some Black children in both reading and racial attitudes.

References

Alexander, K., Entwisle, D., & Olson, L. (2007). Lasting Consequences of the Summer Learning Gap. *American Sociological Review, 72*(2), 167-180.

Allor, J., & McCathren, R. (2004). The efficacy of an early literacy tutoring program implemented by college students. *Learning Disabilities Research & Practice*, *19*(2), 116-129.

Allor, J. H., Cheatham, J. & Al Otaiba, S. (2019). *Friends on the Block.* Richardson, TX: Friends on the Block.

Allor, J. H., Gifford, D. B., Jones, F. G., Otaiba, S. A., Yovanoff, P., Ortiz, M. B., & Cheatham, J. P. (2018). The effects of a text-centered literacy curriculum for students with intellectual disability. *American Journal on Intellectual and Developmental Disabilities*, *123*(5), 474-494.

Allor, J. H., Yovanoff, P., Otaiba, S. A., Ortiz, M. B., & Conner, C. (2020). Evidence for a literacy intervention for students with intellectual and developmental disabilities. *Education and Training in Autism and Developmental Disabilities, 55*(3), 290-302.

Anderson, M. (2020). Y*ou're out of your mind if you think I'm ever going back to school*. The New York Times. Retrieved from https://www.nytimes.com/2020/10/28/opinion/virtual-school-racism.html

Aronson, B., & Laughter, J. (2016). The Theory and Practice of Culturally Relevant Education: A Synthesis of Research Across Content Areas. *Review of Educational Research*, *86*(1), 163–206.

Baker, S. K., Santoro, L. E., Chard, D. J., Fien, H., Park, Y., & Otterstedt, J. (2013). An evaluation of an explicit read aloud intervention taught in whole-classroom formats in first grade. *The Elementary School Journal*, *113*(3), 331-358.

Bell, Y. R., & Clark, T. R. (1998). Culturally relevant reading material as related to comprehension and recall in African American children. *Journal of Black Psychology*, *24*(4), 455-475.

Berninger, V. W., Vaughan, K., Abbott, R. D., Brooks, A., Begayis, K., Curtin, G., ... & Graham, S. (2000). Language-based spelling instruction: Teaching children to make multiple connections between spoken and written words. *Learning Disability Quarterly*, *23*(2), 117-135.

Beverly, B. L., Giles, R. M., & Buck, K. L. (2009). First-grade reading gains following enrichment: Phonics plus decodable texts compared to authentic literature read aloud. *Reading Improvement*, *46*(4), 191-206.

Bishop, R. S. (1990). Windows and mirrors: Children's books and parallel cultures. In *California State University Reading Conference: 14th Annual Conference Proceedings* (pp. 3-12)..

Brown, R., Pressley, M., Van Meter, P., & Schuder, T. (1995). *A quasi-experimental validation of transactional strategies instruction with previously low-achieving, second-grade readers* (Reading Research Report no. 33). Athens, GA: National Reading Research Center.

Bui, Y. N., & Fagan, Y. M. (2013). The effects of an integrated reading comprehension strategy: A culturally responsive teaching approach for fifth-grade students' reading comprehension. *Preventing School Failure: Alternative Education for Children and Youth*, *57*(2), 59-69.

Cartledge, G., Keesey, S., Bennett, J. G., Ramnath, R., & Council III, M. R. (2016). Culturally relevant literature: What matters most to primary-age urban learners. *Reading & Writing Quarterly*, *32*(5), 399-426.

Cervetti, G. N., & Hiebert, E. H. (2015). The sixth pillar of reading instruction: Knowledge development. *The Reading Teacher*, *68*(7), 548-551.
Chambers, T. V. (2009). The "receivement gap": School tracking policies and the fallacy of the " achievement gap.". *The Journal of Negro Education*, 417-431.
Cheatham, J. P., Allor, J. H., & Roberts, J. K. (2014). How does independent practice of multiple-criteria text influence the reading performance and development of second graders?. *Learning Disability Quarterly*, *37*(1), 3-14.
Chu, M. C., & Chen, S. H. (2014). Comparison of the effects of two phonics training programs on L2 word reading. *Psychological Reports*, *114*(1), 272-291.
Clark, L. A., & Watson, D. (1995). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 309-319.
Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, *66*(3), 227-268.
Cox, B. E., McIntosh, K., Reason, R. D., & Terenzini, P. T. (2014). Working with missing data in higher education research: A primer and real-world example. *The Review of Higher Education*, *37*(3), 377-402.
Cummins, J. (2007). Pedagogies for the poor? Realigning reading instruction for low-income students with scientifically based reading research. *Educational Researcher*, *36*(9), 564-572.
Dahir, M. (2019). Between cultural literacy and cultural relevance: A culturally pragmatic approach to reducing the black-white achievement gap. *Handbook of Theory and Research in Cultural Studies and Educatio, 86*(1), 1-19.
de Graaff, S., Bosman, A. M., Hasselman, F., & Verhoeven, L. (2009). Benefits of systematic phonics instruction. *Scientific Studies of Reading*, *13*(4), 318-333.
Duke, N. K., Pearson, P. D., Strachan, S. L., & Billman, A. K. (2011). Essential elements of fostering and teaching reading comprehension. In S. J. Samuels & A. E. Farstrup (Eds.), *What Research Has to Say About Reading Instruction* (4th ed.) (pp. 51-93). Newark, DE: International Reading Association.
Dumont, H., & Ready, D. D. (2020). Do schools reduce or exacerbate inequality? How the associations between student achievement and achievement growth influence our understanding of the role of schooling. *American Educational Research Journal*, *57*(2), 728-774.
Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75*(3), 327–346.
Entwisle, D., Alexander, K., & Olson, L. (1997). *Children, Schools and Inequality*. Boulder, CO: Westview Press.
Fien, H., Smith, J. L., Smolkowski, K., Baker, S. K., Nelson, N. J., & Chaparro, E. (2015). An Examination of the Efficacy of a Multitiered Intervention on Early Reading Outcomes for First Grade Students at Risk for Reading Difficulties. *J Learn Disabil, 48*(6), 602-621.
Garth-McCullough, R. (2008). Untapped cultural support: The influence of culturally bound prior knowledge on comprehension performance. *Reading Horizons: A Journal of Literacy and Language Arts*, *49*(1), 3.
Gay, G. (2010). *Culturally Responsive Teaching: Theory, Research and Practice*. Teachers College Press.
Gershenson, S., & Papageorge, N. (2018). The power of teacher expectations: How racial bias hinders student attainment. *Education Next*, *18*(1), 64-71.

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, *52*, 209-224.
Good, R.H., & Kaminski, R. A. (Eds.) (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.
Hatcher, P. J., Hulme, C., & Snowling, M. J. (2004). Explicit phoneme training combined with phonic reading instruction helps young children at risk of reading failure. *Journal of child Psychology and Psychiatry*, *45*(2), 338-358.
Hayes, D. P., & Grether, J. (1983). The school year and vacations: When do students learn? *Cornell Journal of Social Relations, 17*, 56-71.
Henbest, V. S., & Apel, K. (2017). Effective word reading instruction: What does the evidence tell us? *Communication Disorders Quarterly*, *39*(1), 303-311.
Hiebert, E. & Fisher, C. (2016). A comparison of the effects of two phonetically regular text types on young English learners' literacy. TextProject, Inc.
Hilaski, D. (2020). Addressing the mismatch through culturally responsive literacy instruction. *Journal of Early Childhood Literacy*, *20*(2), 356-384.
Howard, T. C. (2001). Powerful Pedagogy for African American Students: A Case of Four Teachers. *Urban Education*, *36*(2), 179–202.
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.
Hung, M., Smith, W. A., Voss, M. W., Franklin, J. D., Gu, Y., & Bounsanga, J. (2020). Exploring student achievement gaps in school districts across the United States. *Education and Urban Society*, *52*(2), 175-193.
Jackson, T. O. (2011). Developing sociopolitical consciousness at Freedom Schools: Implications for culturally responsive teacher preparation. *Teaching Education*, *22*(3), 277-290.
Jackson, T. O., & Boutte, G. S. (2009). Liberation literature: Positive cultural messages in children's and young adult literature at freedom schools. *Language Arts*, *87*(2), 108-116.
Jackson, T. O., & Howard, T. C. (2014). The continuing legacy of freedom schools as sites of possibility for equity and social justice for Black students. *Western Journal of Black Studies*, *38*(3).
Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading*, *8*(1), 53-85.
Juel, C., & Roper Schneider, D. (1985). The influence of basal readers on first grade reading. *Reading Research Quarterly*, *20*(2), 134-152.
Kaefer, T. (2020). When did you learn it? How background knowledge impacts attention and comprehension in read-aloud activities. *Reading Research Quarterly*, *55*, S173-S183.
Kim, J. S. (2006). Effects of a voluntary summer reading intervention on reading achievement: Results from a randomized field trial. *Educational Evaluation and Policy Analysis*, *28*(4), 335-355.
Kim, J. S., & Quinn, D. M. (2013). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, *83*(3), 386-431.
Koss, M. D. (2015). Diversity in contemporary picture books: A content analysis. *Journal of Children's Literature*, *41*(1), 32-42.

Kramarczuk Voulgarides, C., Fergus, E., & King Thorius, K. A. (2017). Pursuing equity: Disproportionality in special education and the reframing of technical solutions to address systemic inequities. *Review of Research in Education*, *41*(1), 61-87.
Kuhfeld, M. (2019). Surprising new evidence on summer learning loss. *Phi Delta Kappan*, *101*(1), 25-29.
Kuhn, M. R., Schwanenflugel, P. J., Morris, R. D., Morrow, L. M., Woo, D. G., Meisinger, E. B., ... & Stahl, S. A. (2006). Teaching children to become fluent and automatic readers. *Journal of Literacy Research*, *38*(4), 357-387.
Ladson-Billings, G. (1992). Reading between the lines and beyond the pages: A culturally relevant approach to literacy teaching. *Theory into practice*, *31*(4), 312-320.
Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465-491.
Lara-Cinisomo, S., Taylor, D. B., & Medina, A. L. (2020). Summer reading program with benefits for at-risk children: Results from a freedom school program. *Reading & Writing Quarterly*, *36*(3), 211-224.
Lee and Low Books (2018). The diversity gap in children's books. Retrieved from https://i0.wp.com/blog.leeandlow.com/wp-content/uploads/2018/05/Childrens-Books-Infographic-2018.jpg?ssl=1
Lennox, S. (2013). Interactive read-alouds—An avenue for enhancing children's language for thinking and understanding: A review of recent research. *Early Childhood Education Journal*, *41*(5), 381-389.
McClure, E. L., & Fullerton, S. K. (2017). Instructional interactions: Supporting students' reading development through interactive read-alouds of informational texts. *The Reading Teacher*, *71*(1), 51-59.
McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247-288.
Meaney, K, Mickey, K, & editors. (2017). *K-12 reading market survey report 2017*. Simba Information.
Mesmer, H. A. E. (2005). Text decodability and the first-grade reader. *Reading and Writing Quarterly, 21*, 61-86.
Moffatt, L., Heydon, R., & Iannacci, L. (2019). Helping out, signing up and sitting down: The cultural production of "read-alouds" in three kindergarten classrooms. *Journal of Early Childhood Literacy*, *19*(2), 147-174.
Morrison, K. A., Robbins, H. H., & Rose, D. G. (2008). Operationalizing culturally relevant pedagogy: A synthesis of classroom-based research. *Equity & Excellence in Education*, *41*(4), 433-452.
Morrow, L. M., Pressley, M., & Smith, J. K. (1995). *The effect of a literature-based program integrated into literacy and science instruction on achievement, use, and attitudes toward literacy and science* (Reading Research Report no. 37). College Park, MD: National Reading Research Center.
NAEP Reading Report Card. (2019). Retrieved from https://nces.ed.gov/nationsreportcard/
National Center for Education Statistics (2020). *Public school students eligible for free or reduced-price lunch*. Retrieved from https://nces.ed.gov/fastfacts/display.asp?id=898
National Center for Education Statistics (2020). *Characteristics of Public School Teachers.* Retrieved from

https://nces.ed.gov/programs/coe/indicator_clr.asp#:~:text=In%202017%E2%80%9318%2C%20about%2079,1%20percent%20of%20public%20school
National Reading Panel (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Rockville, MD: NICHD Clearinghouse.
Nicholson, T., & Tiru, S. (2019). Preventing a summer slide in reading–the effects of a summer school. *Australian Journal of Learning Difficulties*, *24*(2), 109-130.
Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational researcher*, *41*(3), 93-97.
Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology*, *8*(3), 317-344.
Pericola Case, L., Speece, D. L., Silverman, R., Ritchey, K. D., Schatschneider, C., Cooper, D. H., ... & Jacobs, D. (2010). Validation of a supplemental reading intervention for first-grade children. *Journal of Learning Disabilities*, *43*(5), 402-417.
Petersen, D. B., Allen, M. A., & Spencer, T. D. (2016). Predicting reading difficulty in first grade using dynamic assessment of decoding in early kindergarten: A large-scale longitudinal study. *Journal of Learning Disabilities, 49(2)* 200-215.
Piper, R. E. (2019). Navigating Black Identity Development: The Power of Interactive Multicultural Read Alouds with Elementary-Aged Children. *Education Sciences*, *9*(2), 141.
Puzio, K., Colby, G. T., & Algeo-Nichols, D. (2020). Differentiated Literacy Instruction: Boondoggle or Best Practice?. *Review of Educational Research*, *90*(4), 459-498.
Quinn, D. (2015). Black-White Summer Learning Gaps: Interpreting the Variability of Estimates Across Representations. *Educational Evaluation and Policy Analysis, 37*(1), 50-69. Retrieved November 22, 2020, from http://www.jstor.org/stable/43773486
Quinn, D. M., & Le, Q. T. (2018). Are we trending to more or less between-group achievement inequality over the school year and summer? Comparing across ECLS-K cohorts. *AERA Open*, *4*(4).
Rupley, W. H., Blair, T. R., & Nichols, W. D. (2009). Effective reading instruction for struggling readers: The role of direct/explicit teaching. *Reading & Writing Quarterly*, *25*(2-3), 125-138.
Sampson, R. J., & Sharkey, P. (2008). Neighborhood selection and the social reproduction of concentrated racial inequality. *Demography*, *45*(1), 1-29.
Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). Improving Reading Comprehension in Kindergarten through 3rd Grade: IES Practice Guide. NCEE 2010-4038. *What Works Clearinghouse*.
Slates, S., Alexander, K., Entwisle, D., & Olson, L. (2012). Counteracting summer slide: Social capital resources within socioeconomically disadvantaged families. *Journal of Education for Students Placed at Risk, 17*(3), 165–185.
Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1-26.
Smith, E. P., Atkins, J., & Connell, C. M. (2003). Family, school, and community factors and relationships to racial–ethnic attitudes and academic achievement. *American Journal of Community Psychology*, *32*(1-2), 159-173.
Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.

Souto-Manning, M. (2009). Negotiating culturally responsive pedagogy through multicultural children's literature: Towards critical democratic literacy practices in a first grade classroom. *Journal of Early Childhood Literacy*, *9*(1), 50-74.

Spörer, N., Brunstein, J. C., & Kieschke, U. L. F. (2009). Improving students' reading comprehension skills: Effects of strategy instruction and reciprocal teaching. *Learning andIinstruction*, *19*(3), 272-286.

Stein, M., Johnson, B., & Gutlohn, L. (1999). Analyzing beginning reading programs: The relationship between decoding instruction and text. *Remedial and Special Education*, *20*(5), 275-287.

Suggate, S. P. (2016). A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions. *Journal of learning disabilities*, *49*(1), 77-96.

Taylor, B. M., Pearson, P. D., Clark, K., & Walpole, S. (2000). Effective schools and accomplished teachers: Lessons about primary-grade reading instruction in low-income schools. *The Elementary School Journal*, *101*(2), 121-165.

Teale, W., Paciga, K., & Hoffman, J. (2007). Issues in Urban Literacy: Beginning Reading Instruction in Urban Schools: The Curriculum Gap Ensures a Continuing Achievement Gap. *The Reading Teacher, 61*(4), 344-348.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99*(2), 253–273.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). Test of word reading efficiency. Austin, TX: Pro-Ed.

Torgerson, C., Brooks, G., Gascoine, L., & Higgins, S. (2018). Phonics: reading policy and the evidence of effectiveness from a systematic 'tertiary'review. *Research Papers in Education*, *34*(2), 208-238.

University of Oregon (2020a). 8th edition of dynamic indicators of basic early literacy skills (DIBELS®): Technical manual. Eugene, OR: University of Oregon. Available: https://dibels.uoregon.edu

University of Oregon (2020b). DIBELS® 8th edition administration supplement: Updates to 2020-2021 academic year testing guidance. Eugene, OR: University of Oregon. Retrieved from https://dibels.uoregon.edu/docs/materials/d8/2020-2021-MOY-EOY-DIBELS-Testing-Guidance.pdf

US Census Bureau (2021). American community survey: Median household income, 2015-2019. Retrieved from https://www.census.gov/search-results.html?searchType=web&cssp=SERP&q=household%20income

van Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297-308.

von Hippel, P.T., Workman, J., Downey, D.B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are Schools the Great Equalizer?" *Sociology of Education, 91*, 323–357.

Wang, S., Rubie-Davies, C. M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation*, *24*(3-5), 124-179.

Table 2.1

Child characteristics: Descriptive statistics of pre-tests and outcomes

	N	*Mean/ Percent*	*SD*	*Percent missing*
Child characteristics				
Black	83	0.94	0.24	0
Male	83	0.49	0.50	0
Age	80	6.76	0.98	3.6%
Grade Level (following school year)	83	1.96	0.83	0
Attendance	83	0.90	0.12	0
Socioeconomic proxies				
Zip code, income below nat'l median	83	0.80	0.41	0
High poverty school	83	0.69	0.47	0
Private school	83	0.27	0.44	0
Baseline assessments				
Listening comprehension retell, percent	82	0.68	0.25	1.2%
Word Reading				
Word recognition task, level (out of 12)	51	4.22	3.35	0
Word recognition task, percent	51	0.35	0.28	0
ORF WCPM	29	75.38	32.79	9.4%
ORF, normed percentile	29	0.50	0.58	9.4%
Racial identity scale	82	1.75	1.17	1.2%
Outcome assessments				
Listening comprehension retell, percent	81	0.83	0.20	2.4%
Word Reading				
Word recognition task, level (out of 12)	46	6.78	3.07	9.8%
Word recognition task, percent	46	0.54	0.24	9.8%
ORF WCPM	24	87.26	28.10	25%
ORF, normed percentile	24	0.58	0.26	25%
Racial identity and pride construct	81	2.15	1.12	2.4%

Table 2.2

Paired T-Tests for Gains: Listening Comprehension, Word Reading, and Racial Identity Attitudes

	N	Pre-test *M*	Post-test *M*	Diff	Standardized Difference	*t*
Listening comprehension, percent	80	0.67 (0.23)	0.83 (0.20)	0.16	0.70***	5.05
Word Recognition, levels	46	3.72 (3.07)	6.54 (2.93)	2.83	.92***	14.50
Word Recognition, percent	46	0.31 (0.26)	0.55 (0.24)	0.24	.92***	14.50
ORF WCPM, raw score	22	69.63 (28.49)	90.18 (26.48)	20.55	.72***	5.39
ORF WCPM, percentile	22	0.41 (0.25)	0.60 (0.25)	0.18	.72***	5.32
Racial identity and pride	80	1.74 (1.18)	2.14 (1.12)	0.40	.34*	2.31

Note. Reporting two-tailed t-test results. Standard deviation in parentheses. Statistical significance levels are indicated as ****p<.001, **p<.01,*p<.05.*

Table 2.3

Associations Between Children's Characteristics and Changes: Listening Comprehension, Word Reading, and Racial Identity Attitudes

	(1)	(2)	(3)	(4)	(5)
Listening Comprehension (N = 80)					
Age	0.03				0.03
	(0.02)				(0.02)
Gender		-0.05			-0.06
		(0.04)			(0.04)
Low-income			0.02		0.03
zip code			(0.06)		(0.06)
High poverty				-0.01	-0.03
school				(0.05)	(0.05)
Word Recognition (N = 46)					
Age	0.01				0.01
	(0.02)				(0.02)
Gender		0.02			0.01
		(0.03)			(0.03)
Low-income			-0.08*		-0.08†
zip code			(0.04)		(0.04)
High poverty				0.00	0.01
school				(0.03)	(0.03)
ORF (N=22)					
Age	0.04				0.06†
	(0.04)				(0.03)
Gender		-0.05			-0.10
		(0.07)			(0.06)
Low-income			-0.05		-0.24*
zip code			(0.08)		(0.11)
High poverty				0.09	0.19*
school				(0.07)	(0.07)
Racial Identity (N = 80)					
Age	0.13				0.10
	(0.13)				(0.13)
Gender		0.39			0.35
		(0.25)			(0.26)
Low-income			-0.38		-0.39
zip code			(0.28)		(0.30)
High poverty				-0.03	0.04
school				(0.26)	(0.27)

Note. Robust standard error in parentheses. Statistical significance level indicated as *$p<.05$. Marginal statistical significance level indicated as †$p<.10$.

Appendices

Appendix 1.A Example Decodable Text and Lesson

2.1 DECODABLES 2.0 PROJECT

Animal Sounds

By Julia Lindsey
Illustrated by Meghan Shea

In this text:	
82%	decodable for first grade readers
Focus: Double consonant endings	20% of words in this book feature a double consonant (-zz, -ff, -ss, -ll)
This week's trick words	was, said
Challenge words	what, bee, fly, flowers, live(s), dolphin, snake
Unit 2 Topic Question	"What do animals do?"

Buzz! Buzz! What said buzz?
It is one small bee.
The bee will buzz by the flowers.
Bees live by small flowers and by tall flowers.
Click! Click! What said click?
The dolphin said click. The dolphin will do a flip.
Dolphins swim and flip.
Hiss! Hiss! What said hiss?
It is a snake. The snake sits in the sun.
Snakes sit in the sun on rocks and grass.
Yiff! Yiff! What said yiff?
A red fox said yiff. The fox will run up the hill.
The red fox lives in a small den in the grass.

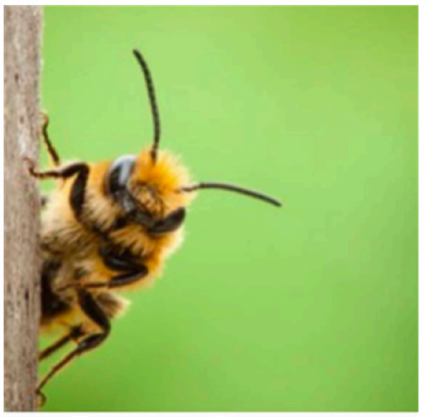
- Introduce this book to the-whole class during a phonics-lesson or before literacy stations.
- Introduce challenge words by analyzing each word (see the-whole group lesson plan).
- Make a unit connection to what animals do and their features.

- Offer this book to read independently during literacy stations or another independent reading time.
- Tell students to read 2-3 times. Encourage students to discuss-the book with a partner
- Students can also underline,highlight, or rewrite words with the phonics focus from the text as they say them aloud after reading.

- For students who are practicing applying phonics to reading, use this book in place of your typical small group reading 1 time during theweek.
- This book should be used withstudents during Focus U2W1and Fundations U4W2

- At the end of the week during which you used this book, send a copy home with each student to read 3 times.
- Have students record their reading with the Decodables At Home chart

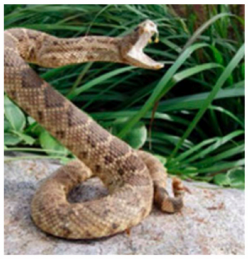
Animal Sounds

 Buzz! Buzz! What said buzz?	 It is one small bee.
 The bee will buzz by the flowers.	 The bees live by small flowers and tall flowers.
 Click! Click! What said click?	 The dolphin said click. The dolphin will do a flip.

Dolphins swim and flip.

Hiss! Hiss! What said hiss?

It is a snake! The snake sits in the sun.



Snakes sit in the sun on rocks and grass.

Yiff! Yiff! What said yiff?

A red fox said yiff. The fox will run up the hill.

The red fox lives in a small den in the grass.

Appendix 1.B: Text Characteristics

Appendix 1.B Table 1

Quantitative Criteria by Text

Text			Word decodability		Word frequency				Syllables	Morphemes	Word Understandability (of non-decodable words)		
Title	Total words	Mean sentence length (words)	Percentage of words with learned grapheme-phoneme correspondence or learned HFWs (decodability)	Percentage of words ending in highly frequent rimes	Percentage of words in top 100-word list (Fry high frequency words)	Percentage of words that are learned high frequency words (Fundations)	Type-token ratio	Percentage of singlets that are decodable or high frequency words	Total instances of multisyllabic words	Average morphemes per word	Percentage of concrete words (of non-decodable words)	Percentage of imageable words (of non-decodable words)	Percentage of familiar or content vocabulary words (of non-decodable words)
Fun Families	51	8.5	92.16%	33.33%	66.67%	52.94%	0.57	80.95%	2	1.05	80.00%	80.00%	80.00%
Play Ball	50	3.57	85.42%	40.00%	27.08 %	22.91%	0.44	87.50 %	3	1.10	85.42%	71.43%	100%
Can We Get a Pet?	66	9.43	90.77%	36.92%	56.92%	33.84%	0.59	77.78%	1	1.03	50.00%	83.33%	100%
Kick It	65	3.82	90.77%	47.69%	40.00%	30.77%	0.48	94.44%	0	1.18	100%	100%	100%
Shop with Mom	85	7.08	93.84%	42.68%	30.48%	25.61%	0.45	85.71	2	1.19	100%	100%	100%
Recess	70	5.83	92.75%	34.78%	27.54%	21.74%	0.40	87.50%	8	1.20	100%	100%	100%
First Grade	69	5.31	82.60%	24.64%	44.93%	27.54%	0.59	95.00%	2	1.16	75.00%	50.00%	100%
Boston in Fall	79	6.58	81.01%	46.83%	45.57%	30.38%	0.49	79.16%	12	1.18	100%	100%	100%
Snack Jobs	74	6.73	87.83%	33.78%	43.24%	28.38%	0.57	79.16%	5	1.24	77.78%	77.78%	100%
Bus Ride	75	5.36	93.24%	22.97%	60.81%	39.19%	0.45	89.74%	3	1.05	60.00%	100%	100%
My Block	51	5.67	90.20%	23.53%	50.98%	35.29%	0.75	82.14%	2	1.17	40.00%	80.00%	100%
I am a Leader	89	5.93	84.27%	28.09%	66.29%	48.31%	0.51	62.96%	9	1.10	57.14%	71.43%	92.86%
Leader Quiz	56	7.00	85.71%	17.86%	67.87%	51.79%	0.46	64.28%	6	1.05	47.50%	75.00%	100%
Dear Librarian	81	5.79	81.01%	18.99%	50.63%	40.51%	0.48	61.11%	3	1.14	70.59%	70.59%	100%
Rick and the Dock	112	8.62	86.60%	45.53%	53.57%	45.54%	0.34	81.81%	5	1.15	0%	33.33%	100%
The Boston Public Market	125	8.33	83.32%	27.87%	53.28%	31.97%	0.37	100%	20	1.11	70.83%	83.33%	91.62%
Animal Sounds	97	4.04	85.57%	37.11%	48.45%	36.08%	0.42	80.00%	6	1.04	100%	100%	100%
Ducks in Boston	103	9.36	80.77%	52.43%	45.63%	23.30%	0.39	100%	16	1.30	50.00%	60.00%	100%
A Mouth for a Meal	98	6.12	80.61%	35.71%	36.73%	29.59%	0.57	90.90%	6	1.23	81.81%	100%	95.45%
Rainforest Life	113	7.53	80.00%	57.39%	51.30%	38.26%	0.40	86.36%	17	1.31	92.31%	92.31%	100%
Animal Babies	178	7.12	80.33%	34.83%	48.88%	37.08%	0.31	83.33%	19	1.16	80.00%	91.43%	100%
Is it a Chipmunk?	160	5.40	90.63%	43.75%	61.88%	44.38%	0.36	68.42%	3	1.08	73.33%	86.67%	100%
Sea Turtle Babies	148	7.05	81.76%	33.78%	54.73%	40.54%	0.41	100%	11	1.11	60.00%	73.33%	100%
Dr. Kim's Sea Turtles	165	10.31	81.32%	19.87%	54.82%	36.14%	0.38	100%	10	1.36	69.44%	88.89%	100%
Back at the Boston Public Market	197	8.57	80.32%	15.73%	53.81%	39.08%	0.35	90%	25	1.26	75.00%	91.67%	100%
Amber's Birthday	245	6.12	86.89%	24.27%	51.44%	37.45%	0.33	90.90%	33	1.30	26.67%	26.67%	93.33 %
A Cocoa Farm	189	9.45	84.48%	34.04%	50.00%	34.04%	0.47	92.50%	28	1.38	74.07%	92.00%	100%
What's in a Pancake?	242	8.96	81.58%	37.19%	52.48%	34.71%	0.38	94.73%	41	1.42	61.90%	71.43%	95.24%
For Lunch	297	7.82	90.68%	37.54%	46.42%	38.91%	0.36	86.67%	39	1.40	76.92%	73.08%	88.46%
Firefighters	283	9.43	83.58%	28.37%	60.99%	50.71%	0.39	96.00%	44	1.19	88.64%	93.81%	100%
Paul's Budget	300	8.57	97.99%	30.87%	46.64%	36.91%	0.41	92.11%	15	1.24	50.00%	16.67%	100%
The Store Down the Street	150	N/A (poem)	92.81%	24.18%	64.71%	53.59%	0.44	92.68%	12	1.12	50.00%	50.00%	100%
Genius Gia Hears Boston	358	7.46	94.69%	24.58%	55.58%	49.16%	0.35	94.54%	19	1.18	73.68%	78.94%	100%
Genius Gia Makes a Sound	194	9.24	90.96%	35.11%	51.06%	39.89%	0.46	93.93%	23	1.13	94.11%	100%	100%
Genius Gia and the Solar Fountain	298	8.76	93.22%	29.15%	45.42%	41.69%	0.40	93.88%	45	1.22	45.00%	55.00%	60.00%
Genius Gia Goes to the Zoo	305	7.44	95.69%	39.73%	47.68%	42.38%	0.40	98.41%	58	1.25	76.92%	84.61%	100%
Genius Gia Makes a Kaleidoscope	239	8.85	96.10%	28.71%	53.67%	40.69%	0.44	95.74%	36	1.19	88.89%	100%	100%
Genius Gia and the Safe Streets	337	10.21	93.73%	31.94%	48.96%	36.71%	0.44	91.78%	62	1.31	60.00%	65.00%	100%
Genius Gia and the Kid Creators	232	7.80	92.67%	32.76%	53.88%	43.53%	0.45	85.71%	32	1.25	17.64%	29.41%	52.94%
Genius Gia Stays Home	232	9.28	98.25%	32.17%	50.43%	43.91%	0.44	92.45%	35	1.36	50.00%	50.00%	75.00%
Mean (sd)	*158.95 (91.43)*	*7.40 (1.75)*	*87.90% (5.59%)*	*33.19% (9.25%)*	*50.51% (9.36%)*	*37.86% (8.08%)*	*0.44 (0.08)*	*87.51% (13.25%)*	*17.95 (16.44)*	*1.20 (0.10)*	*68.39% (22.95%)*	*75.64% (22.31%)*	*95.66% (10.51%)*

Appendix 1.B Table 2

Qualitative Criteria by Text

Title	Synopsis	Content Connection	Potential background knowledge connection	Persons Represented
Fun Families	Three children show their family engaged in an activity.	Families and communities unit in K	Life as a child in different families	Black/African American, Asian American, LGBTQ+
Play Ball	A child plays fetch with dad	Families and communities unit in K	Playing, city neighborhood	Black/African American
Can We Get a Pet?	Classroom wants a class pet	Families and communities unit in K	BPS first grade classroom	Ethnically and racially diverse children
Kick It	Children's soccer game	Families and communities unit in K	Soccer field	Ethnically and racially diverse children
Shop with Mom	Children shopping with mother	Families and communities unit in K	City neighborhood, bus	Black/African American
Recess	Children inviting others to play at recess	Families and communities unit in K	Playground	Ethnically and racially diverse children
First Grade	Children on the first day of school	Families and communities unit in K	BPS first grade classroom	Ethnically and racially diverse children
Boston in Fall	Boston scenes in fall	Families and communities unit in K	Typical Boston fall activities and sports teams	Ethnically and racially diverse children
Snack Jobs	Children eat and help out during snack	Helping out in the classroom	BPS classroom with typical snack time rituals	Ethnically and racially diverse children
Bus Ride	Children see their neighborhood from a bus	Unit text *The Last Stop on Market Street*	Riding a bus through Boston	Ethnically and racially diverse persons, persons in Hijab
My Block	Children help with tasks on their block	Helping out in the neighborhood	City neighborhood	Ethnically and racially diverse characters
I am a Leader	Learning about leader character traits by comparing to familiar leaders	Learning about leaders	Photos of familiar leaders to children in USA, Boston	Barack Obama, Marty Walsh, etc.
Leader Quiz	Learning about leader character traits	Learning about leaders	Photos in familiar scenes	Ethnically and racially diverse characters
Dear Librarian	Asking a librarian for help	Community helpers, letter writing	School library	Black male librarian
Rick and the Dock	A boy cleans up a local dock	Helping out in the neighborhood	City neighborhood near harbor	Black/African American mother and son
The Boston Public Market	Investigating stalls at the Boston Public Market	Community features	Real stalls at the Boston Public Market	Black/African American shop keeper
Animal Sounds	Animals make different noises	Animal features	Familiar animals (bees, dolphins)	N/A
Ducks in Boston	Ducks live in specific habitats	Animal habitats	Duck habitats in Boston parks	N/A
A Mouth for a Meal	Animals have mouths for purposes	Animal features help them survive	Familiar animals (cats, ducks)	N/A
Rainforest Life	The rainforest is one habitat	Animal habitats	Potentially none beyond science content	N/A
Animal Babies	Animals start out as babies	Animal baby survival	Familiar animals (dogs, cats)	N/A
Is it a Chipmunk?	Identifying a chipmunk from other animals	Animal features	Chipmunks in urban parks	N/A
Sea Turtle Babies	Sea turtle babies first moments of life	Animal baby survival	Potentially none beyond science content	N/A
Dr. Kim's Sea Turtles	Scientists help sea turtles	Animal baby survival	Potentially none beyond science content	Real-life female scientist
Back at the Boston Public Market	A family goes shopping at the market	Markets and shopping	Real stalls at the Boston Public Market	Latino family
Amber's Birthday	A girl opens birthday presents	Needs and wants	Birthdays	Black/African American family
A Cocoa Farm	How chocolate is made	Where resources come from	Chocolate	N/A
What's in a Pancake?	How to make a pancake and where the ingredients come from	Where resources come from, how-to text writing	Breakfast foods	N/A
For Lunch	Where lunch food comes from	Where resources come from	Lunchroom at a typical school	Ethnically and racially diverse students
Firefighters	What do firefighters do	Jobs and services	Firefighters	Gender, ethnically, and racially diverse firefighters
Paul's Budget	A boy learns how to budget	Making consumer choices	City neighborhood	Indian American aunt and child
The Store Down the Street	A poem about a local store	Consumer choices impact the community, poetry writing	A real local store in Boston	Ethnically and racially diverse characters
Genius Gia Hears Boston	Gia finds the source of different sounds	Sounds as vibrations	City neighborhood	Latino family
Genius Gia Makes a Sound	Gia makes a guitar	Sounds as vibrations, how-to text writing	Typical apartment	Latino family
Genius Gia and the Solar Fountain	Gia learns about solar power	People use light, light as waves	City neighborhood, Spanish language	Latino family, Spanish speakers
Genius Gia Goes to the Zoo	Gia hears different animals at the zoo	Animals use sound	Franklin Park Zoo	Ethnically and racially diverse characters; Black male teacher
Genius Gia Makes a Kaleidoscope	Gia makes a kaleidoscope	Light can change, how-to text writing	Typical apartment	Latino family
Genius Gia and the Safe Streets	Gia learns about Garrett Morgan	People use light, inventors make a difference	City neighborhood, Spanish language	Latino family, Spanish speakers, Black/African American inventor
Genius Gia and the Kid Creators	Gia learns that kids can be inventors	Inventors make a difference	City neighborhood	Latino family, Spanish speakers; young Black female inventors
Genius Gia Stays Home	Gia supports her community during a COVID lockdown	N/A	COVID-19 pandemic lockdowns and other restrictions	Latino family

Appendix 1.B Table 3
Select Criteria by Text Set

Text set	Unit Topic	Texts linked to	Total words	Number of singlets	Type-token ratio
1	Living in Boston	Assumed background knowledge	535	76	0.33
2	Community	Social studies unit	663	76	0.28
3	Animals	Science unit	1,061	76	0.22
4	Resources, needs/wants	Social studies unit	1,893	152	0.22
5	Sound and light	Science unit	2,197	171	0.20

Appendix 1.C Comparison to District

Appendix 1.C Table 1
Demographic Data: BPS Schools and Sample Schools (SE)

	Sample	BPS	Standardized Difference
Male %	52.14	52.62	0.10
	(3.66)	(5.32)	
IEP %	21.94	23.45	0.13
	(6.48)	(15.36)	
ELL %	36.54	37.05	0.02
	(13.90)	(19.00)	
Economically disadvantaged %	60.96	58.51	0.15
	(16.42)	(16.27)	
Race and ethnicity categories			
Black %	31.32	28.69	0.14
	(17.31)	(20.78)	
Asian %	7.22	5.39	0.18
	(10.18)	(9.83)	
Hispanic %	40.04	44.71	0.24
	(17.16)	(22.32)	
Other %	4.70	4.23	0.21
	(1.99)	(2.31)	

Appendix 1.D: Robustness Checks

I varied multiple aspects of my analytic approach to ensure results are robust to modeling decisions. Results from these analyses (as described in the Sensitivity Analyses section) are in Table 1. Point estimates and effect sizes remain similar across approaches.

Appendix 1.D Table 1
Robustness Check: Estimated Treatment (Standard Error) Impact on Children's Phonics Assessment with Varied Modeling Approaches

	All district schools	Correct test dates	Dummy variable	Three-level model
Treatment	-0.06	-0.02	-0.03	-0.02
	(0.09)	(0.05)	(0.06)	(0.06)
Effect size	-0.07	-0.05	0.05	-0.04
N	1,885	1,251	1,471	1,391

Note. Standard error in parentheses. Model 1 shows multilevel linear regression comparing treatment schools to all available district schools. Model 2 shows multilevel linear regression comparing students only with tests within the district-mandated testing periods. Model 3 multilevel linear regression with a dummy variable and mean substitution for missing predictor and outcome data. Model 4 shows a three-level multilevel linear regression with random intercepts for both schools and classrooms.

Appendix 1.E: Fidelity Tools

Appendix 1.E Table 1
Observational Tool

Question	Answer Options
Briefly write about the posters/visuals you see that connect to Focus and Fundations.	
What unit and week of Fundations is the teacher on?	
What unit and week of FOF is the teacher on?	
How many students are present?	
In what setting(s) were Decodables 2.0 were used during your observation?	Whole group Small group Independent reading
During your observation, Decodable 2.0 texts are clearly available for students in classroom library, literacy station, or student's individual book sets.	Y/N
Students are reading Decodable 2.0 texts independently during stations.	Y/N
Teacher explicitly instructs or reviews phonics matched to Decodable 2.0 text concepts prior to reading.	Y/N
Teacher explicitly connects Decodable 2.0 to unit topics, weekly questions, or science/social studies content.	Y/N
If the teacher introduced/reviewed a trick word, how did she introduce review?	With word structure By "sight" No
Teacher prompts students with skills, like using a particular phonics letter-sound relationship(s) to decode a word (when appropriate).	Yes, often and responsively Yes, some No
Teacher prompts students with strategies, like tap or sound out words.	Yes, often and responsively Yes, some No
Students spend the majority of time reading.	Y/N
The focus of the lesson is on using phonics knowledge while reading.	Yes No, lesson focuses on something else Not really a lesson

	Not observed
Students are observed using tapping or sounding out to read words in small group or independent reading.	Y/A
Write about your general observations of the teaching.	Open-ended
Listen to students reading. Describe students reading.	Open-ended

Appendix 1.E Table 2

Teacher Survey

Question	Answer Options
Which Decodable 2.0 texts have you used this year?	All text titles available
Which 2 weeks will you referring to in your answers?	Week options
Have you used a Decodables 2.0 text in your classroom in the past 2 weeks?	Y / N
Did you use a Decodable 2.0 during both of the last 2 weeks of instruction?	Y / N
About how many students in your class read a Decodable 2.0 text in the last 2 weeks?	• All students • More than half of class • Less than half of class
In the last 2 weeks, how did you use a Decodable 2.0 text?	• Small group • Whole group • Independent reading • Home reading
How did you primarily decide to use a Decodable 2.0 in the past two weeks?	• I'm using the texts because I said I would for this study. • I chose to use texts that supported my phonics lessons. • I chose Decodables 2.0 texts that fit my students' needs (in any setting). • I needed them to use as a station activity.
How did you primarily decide which students would have a small group with you using a Decodable 2.0 in the past two weeks?	• I used Decodables 2.0 with my "lowest" ability reading group. • I used Decodables 2.0 with my "highest" ability reading group. • I used Decodables 2.0 with my "middle" ability reading group. • I used Decodables 2.0 with all my reading groups. • I used Decodables 2.0 with a different group than my typical reading groups, based on phonics knowledge.
What data did you use to figure out which texts to use with each small group in the past two weeks?	Open-ended
Describe how you think Decodables 2.0 do and/or do not support your students' reading development.	Open-ended
In the past two weeks, how engaged do you think students are when reading Decodables 2.0?	Rate 1 to 5 (very engaged)

Appendix 1.E Table 3

Composite Fidelity Tool

Tool	Item	Options	Score
Observation	During your observation, Decodable 2.0 texts are clearly available for students in classroom library, literacy station, or student's individual book sets.	Yes	1
		No	0
Observation	Students are reading Decodable 2.0 texts independently during stations.	Yes	1
		No	0
Observation	Teacher explicitly instructs or reviews phonics matched to Decodable 2.0 text concepts prior to reading.	Yes	1
		No	0
Observation	Teacher explicitly connects Decodable 2.0 to unit topics, weekly questions, or science/social studies content.	Yes	1
		No	0
Observation	If the teacher introduced/reviewed a trick word, how did she introduce review?	Yes	1
		No	0
Observation	Teacher prompts students with skills, like using a particular phonics letter-sound relationship(s) to decode a word (when appropriate).	Often and responsively	2
		Sometimes	1
		No	0
Observation	Teacher prompts students with strategies, like tap or sound out words.	Often and responsively	2
		Sometimes	1
		No	0
Observation	Students spend the majority of time reading.	Yes	1
		No	0
Observation	The focus of the lesson is on using phonics knowledge while reading.	Yes	1
		No	0
Observation	Students are observed using tapping or sounding out to read words in small group or independent reading.	Yes	1
		No	0
Observation	Write about your general observations of the teaching.		N/A
Observation	Listen to students reading. Describe students reading.		N/A

Survey	Which Decodables 2.0 have you used this year?	Drop down menu	1 = 10-20 texts
Survey	Have you used a Decodable 2.0 in the past 2 weeks?	Yes	1
		No	0
Survey	How many children in your class read a Decodable 2.0 in the last 2 weeks?	All	1
		More than half	1
		Less than half	0
		None	0

Appendix 2.A: Racial Identity Scale

1. Do you know what the word race means? Do you know what your race is?
    a. No answer (0)
    b. No to one or both (1)
    c. Yes, to both with support from parent (2)
    d. Yes, to both (3)
2. Do you like being your race?
    a. No answer (0)
    b. No (1)
    c. Maybe (2)
    d. Yes (3)
3. Should people be proud of their race?
    a. No answer (0)
    b. No (1)
    c. Maybe (2)
    d. Yes (3)

Appendix 2.B: Robustness Checks

Due to relatively high levels of missing data on the ORF post-test assignment and association of age with missingness on this outcome, I conducted robustness checks on models addressing Research Question 2 to check my findings across different missing variable assumptions. I investigated gains on the ORF measure using multiple imputation, first imputing only the predictors and then imputing both predicators and outcomes. As in the complete case analysis, children's characteristics (age, gender, and socioeconomic proxies) were not statistically significantly associated with changes in oral reading fluency in isolation. Point estimates on each predicator in isolation were relatively stable across all three manners of dealing with missing data. When controlling for all covariates, the statistical significance of children's characteristics differed and was not stable, likely due to the low statistical power and potential non-validity of the socioeconomic proxies. These results provide some additional support for the model in the main text as they demonstrate some level of stability across estimates.

Appendix 2.B Table 1

Robustness Check A: Associations Between Children's Characteristics and Gains in Oral Reading Fluency, Multiple Imputation for Predictors

	(1)	(2)	(3)	(4)	(5)
ORF, pretest	0.55*	0.57*	0.52†	0.50†	0.55*
	(0.26)	(0.26)	(0.26)	(0.26)	(0.24)
Male	-0.08				-0.14†
	(0.08)				(0.08)
Age		0.06			0.09†
		(0.05)			(0.05)
Low Income			-0.09		-0.30*
zip code			(0.09)		(0.13)
High poverty				0.06	0.20*
school				(0.08)	(0.08)
Constant	0.39**	-0.09	0.43**	0.33*	-0.12
	(0.12)	(0.38)	(0.15)	(0.12)	(0.34)
Observations	24	24	24	24	24

Note. Robust standard error in parentheses. Statistical significance level indicated as ***$p<.001$, **$p<.01$, *$p<.05$. Marginal statistical significance level indicated as $^{\dagger}p<.10$.

Appendix 1.B Table 2

Robustness Check B: Associations Between Children's Characteristics and Gains in Oral Reading Fluency, Multiple Imputation for Predictors and Outcomes

	(1)	(2)	(3)	(4)
ORF, pretest	0.78***	0.68***	0.68***	0.83***
	(0.17)	(0.18)	(0.18)	(0.16)
Age	0.09†			0.10*
	(0.05)			(0.05)
Male				-0.11
				(0.08)
Low Income		-0.05		-0.07
zip code		(0.07)		(0.08)
High poverty			0.03	0.04
			(0.06)	(0.06)
Constant	-0.41	0.28*	0.22*	-0.44
	(0.37)	(0.10)	(0.09)	(0.37)
Observations	83	83	83	83

Note. Robust standard error in parentheses. Statistical significance level indicated as ***$p<.001$, **$p<.01$, *$p<.05$. Marginal statistical significance level indicated as †$p<.10$.